ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

КАФЕДРА ВТ

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №1

«Введение в R. Методы первичного разведочного анализа данных в R» по дисциплине «Компьютерные технологии анализа и обработки данных»

Выполнили: студенты

гр. АММ2-24

Атласюк Игорь Романович

Ириков Евгений Алексеевич

Проверил: к.т.н., доцент Кафедры ВТ Альсова Ольга Константиновна

Новосибирск 2024

## Содержание

[Постановка задачи 3](#_bookmark0)

[Ход работы 4](#_bookmark1)

[Заключение 18](#_bookmark2)

[Приложение 19](#_bookmark3)

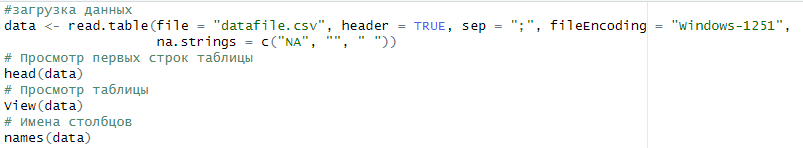
# **Постановка задачи**

Получить базовые навыки работы в среде R. Изучить средства R для проведения первичного разведочного анализа данных (методы визуализации, описательной статистики, корреляционного анализа данных) на примере решения конкретной задачи ИАД (интеллектуального анализа данных).

# **Ход работы**

Для открытия файла и базовыми операциями над ним были использованы такие команды как:

* read.table - считывает файл в формате таблицы и создает из него фрейм данных;
* head - просмотр первых строк таблицы;
* view - просмотр таблицы в отдельном окне;
* names – просмотр имен столбцов.

Рисунок 1. Загрузка данных

Для загрузки данных используется команда read.table с атрибутами file равное названию файла, кодировкой windows-1251, и na.strings чтобы удалять пустые значения.

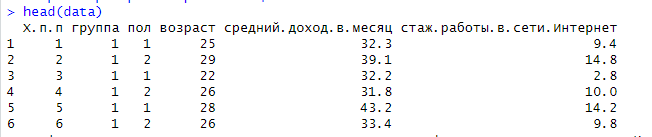


Рисунок 2. Результат работы head

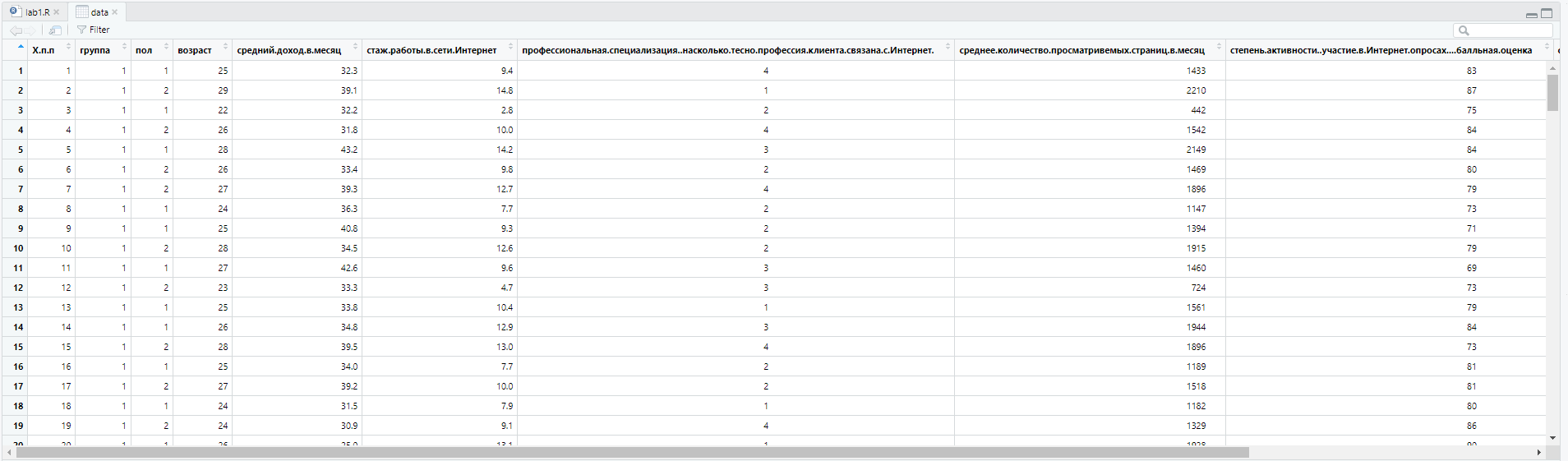


Рисунок 3. Результат работы view

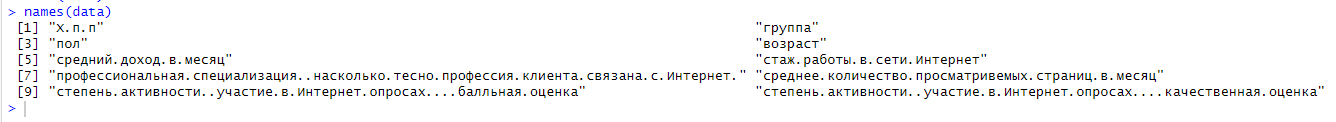


Рисунок 4. Результат работы names

Для обращения к определенному столбцу используется data$’название столбца’, а для написания условий используется команда subset.

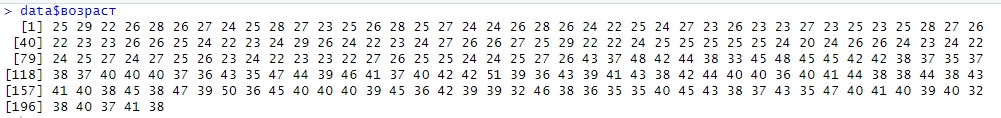


Рисунок 5. Результат работы data$возраст

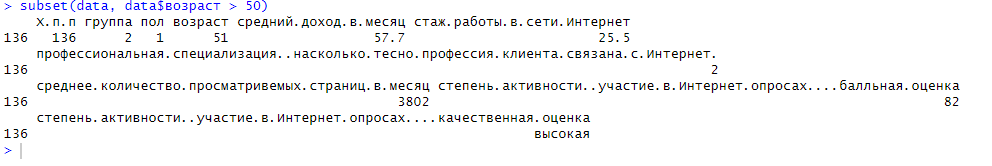


Рисунок 6. Результат работы subset(data, data$возраст > 50)

Далее для просмотра структуры данных, расчета основных статистических характеристик (минимальное, максимальное, среднее значение, стандартное отклонение, первая и третья квартили, медиана, мода, асимметрия, эксцесс) использовались команды описанные ниже на рисунке 7.

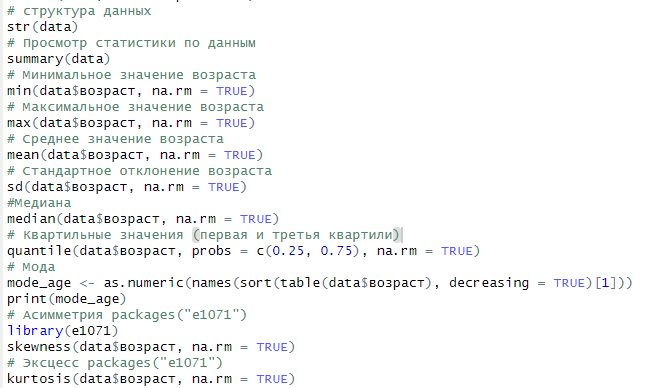


Рисунок 7. Команды для описания основных статистических характеристик

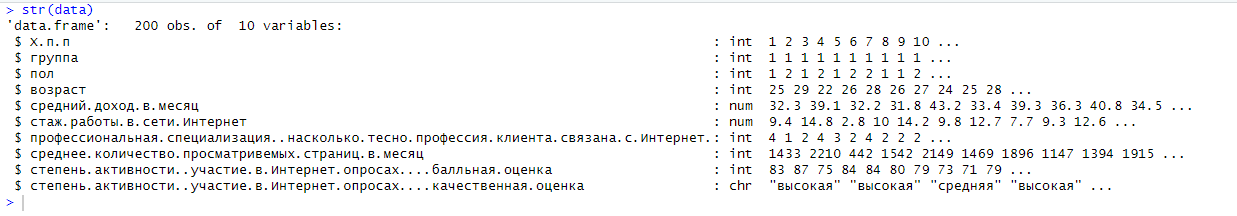


Рисунок 8. Структура данных

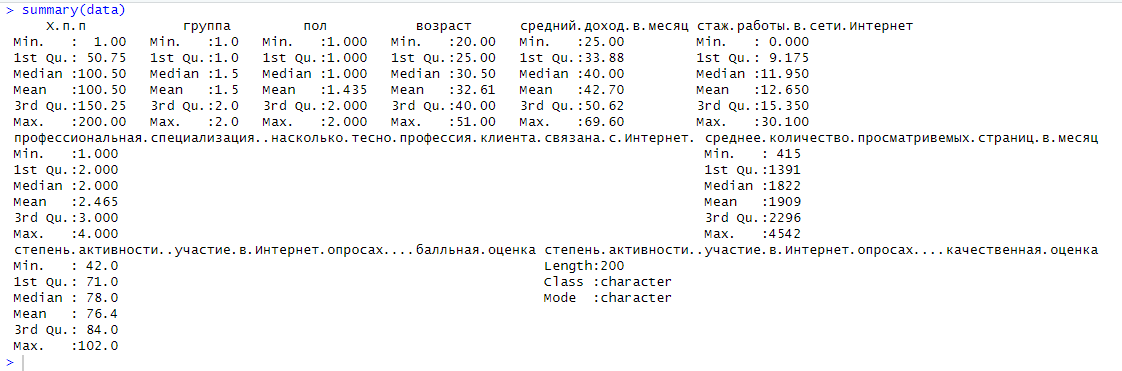


Рисунок 9. Статистика данных

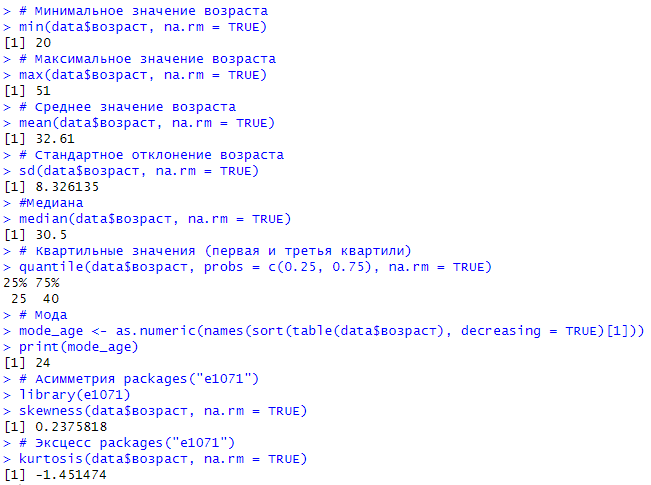


Рисунок 10. Значения по возрасту

На рисунке 10 показаны результаты команд, на основе которых можно сделать следующие выводы:

* Возрастная выборка достаточно вариативна (разброс возрастов значителен).
* Большинство возрастов сосредоточено в диапазоне от 25 до 40 лет, с центром около 30–33 лет.
* Распределение слегка смещено вправо, с присутствием более старших возрастов, но в целом распределение близко к симметричному.
* Мода в 24 года означает, что возраст 24 встречается чаще всего, хотя средний возраст выше.
* Эксцесс показывает, что распределение несколько уплощённое, что может говорить о меньшем числе крайних возрастов (молодых или старых участников).

Далее по заданию нужно провести графический анализ данных и построить следующие диаграммы:

* диаграмму рассеяния по двум количественным признакам;
* радиальную диаграмму по качественному признаку;
* категориальную радиальную диаграмму по одному из качественных признаков в зависимости от групповой переменной (например, от пола, номера группы, наличия кредита);
* категориальную столбиковую диаграмму по одному из количественных признаков в зависимости от групповой переменной (например, от пола, номера группы, наличия кредита и т. п.);
* диаграмму размаха для одного из количественных признаков в зависимости от групповой переменной (например, от пола, номера группы, наличия кредита и т. п.);
* гистограммы для количественных признаков в одном графическом пространстве;
* матричный график для количественных переменных.

Ниже на рисунках 11 и 12 показан код для построения диаграмм.

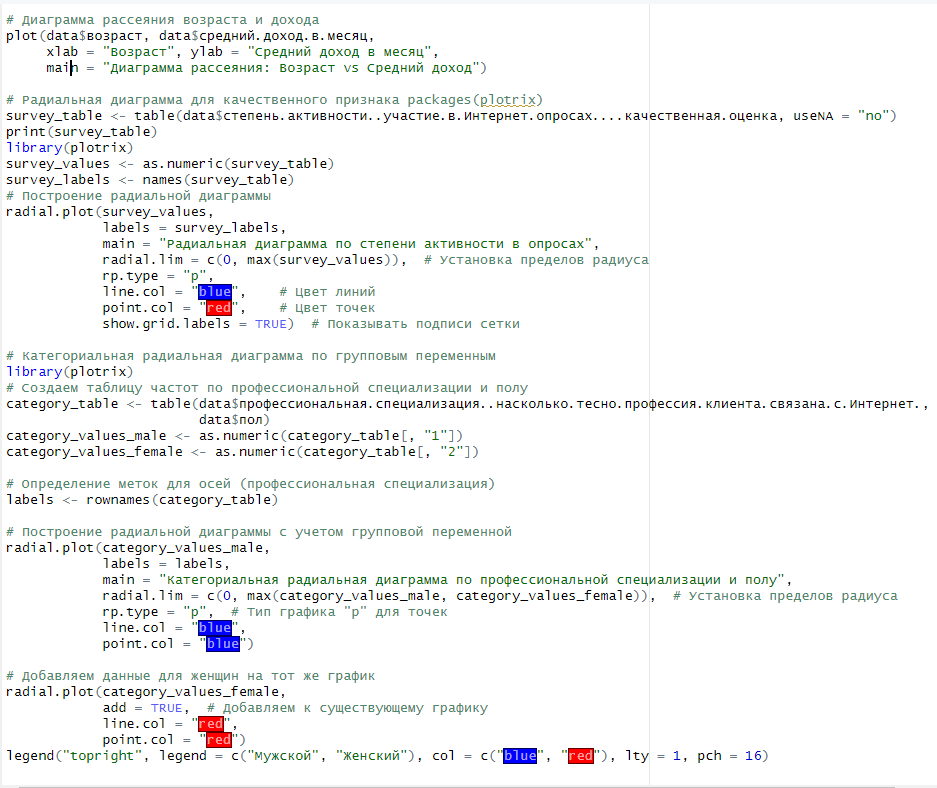


Рисунок 11. Описание диаграмм

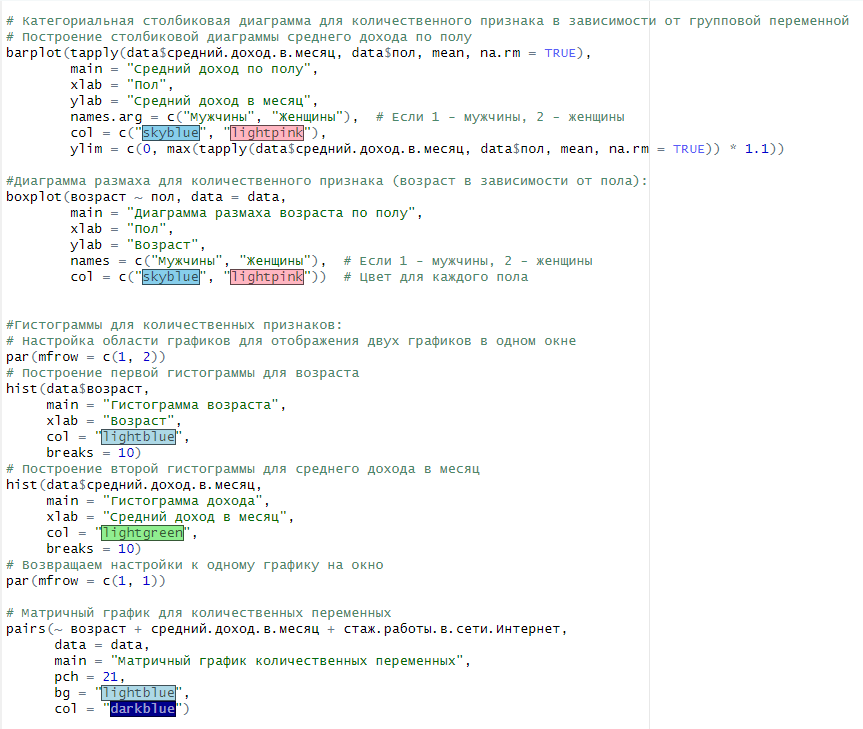


Рисунок 12. Описание диаграмм

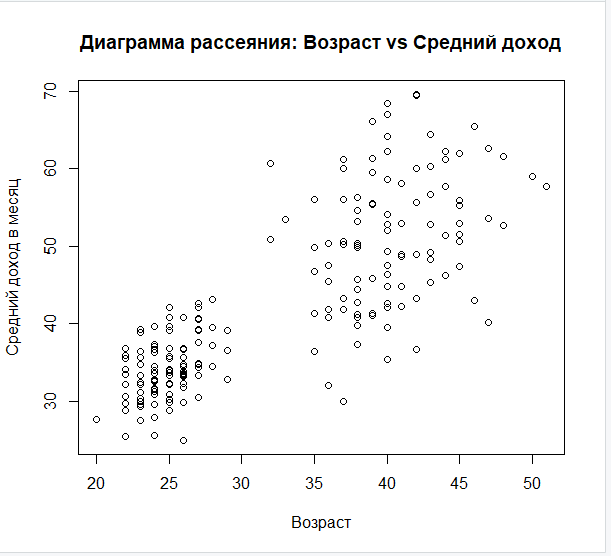


Рисунок 13. Диаграмма рассеяния



Рисунок 14. Радиальная диаграмма

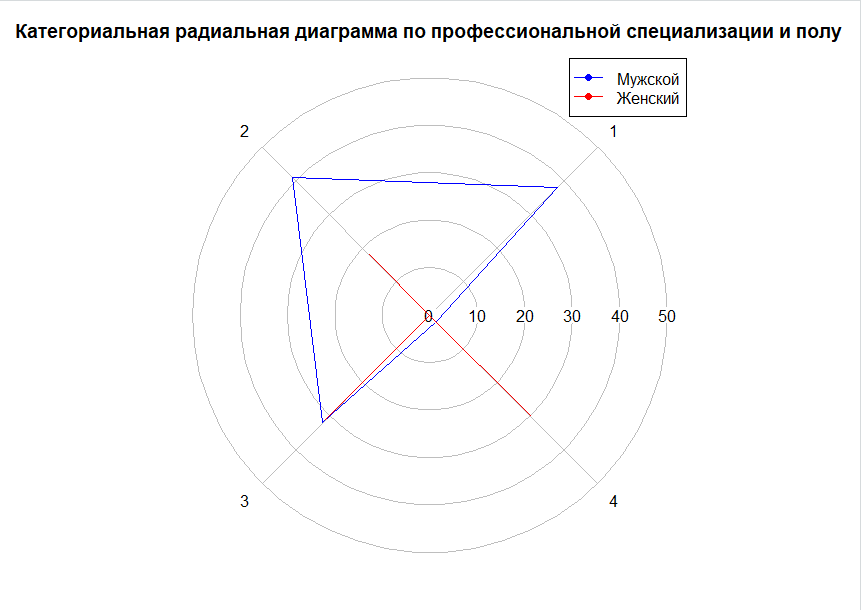


Рисунок 15. Категориальная радиальная диаграмма

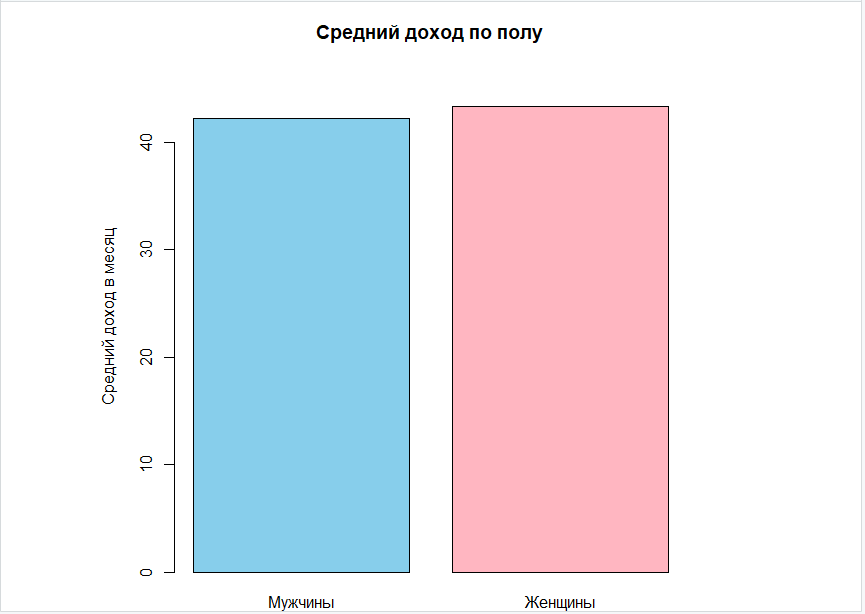


Рисунок 16. Категориальная столбиковая диаграмм

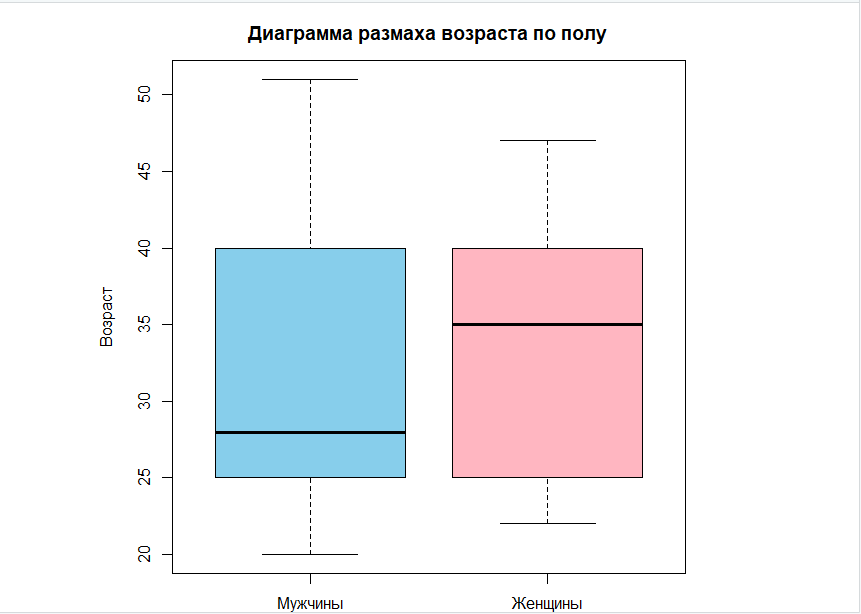


Рисунок 17. Диаграмма размаха

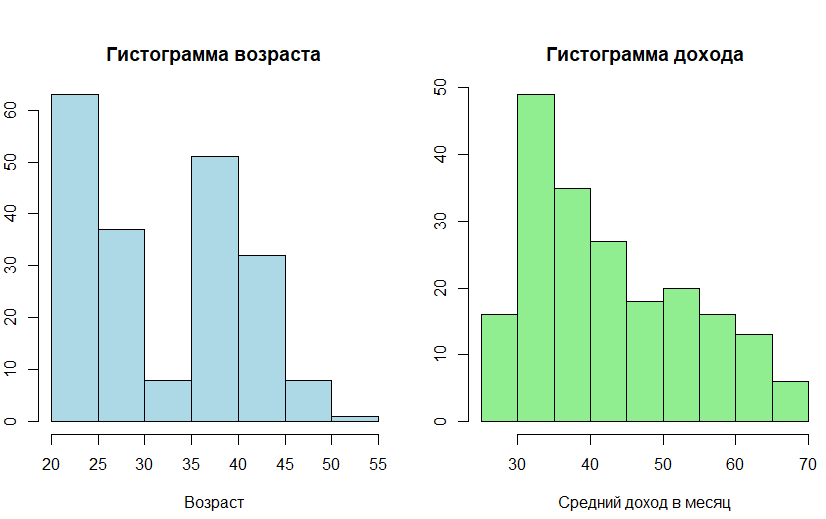


Рисунок 18. Гистограммы для количественных признаков

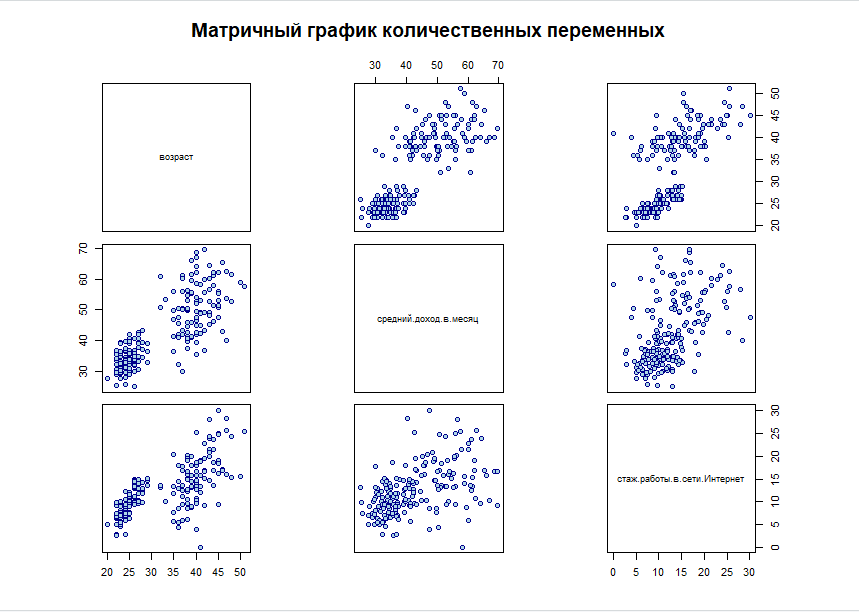


Рисунок 19. Матричный график для количественных переменных

На основе диаграмм можно сделать некоторые выводы, средний доход не зависит от пола (они почти равны), средний доход в месяц 35–45, больше всего опрошенных меньше 30 лет.

Далее был проведен корреляционный анализ.

1. χ2 (Chi-квадрат) и Фишера для первой и второй групп.

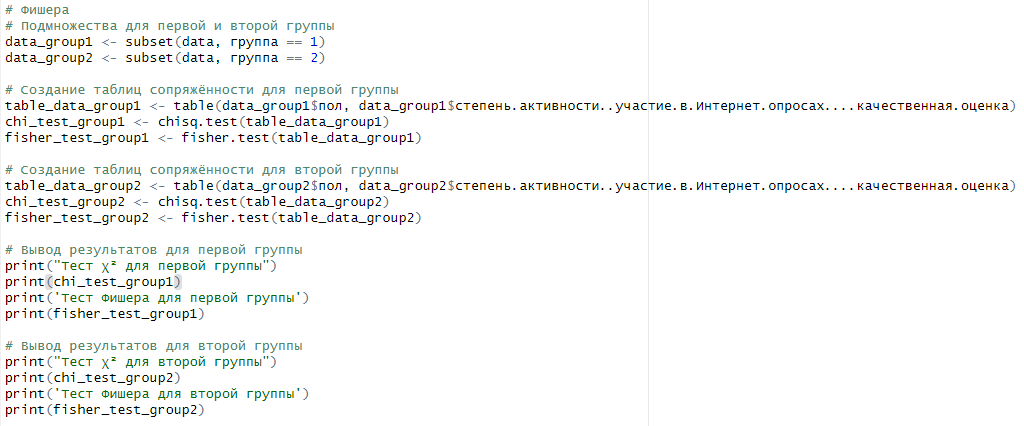


Рисунок 20. χ2 и Фишер

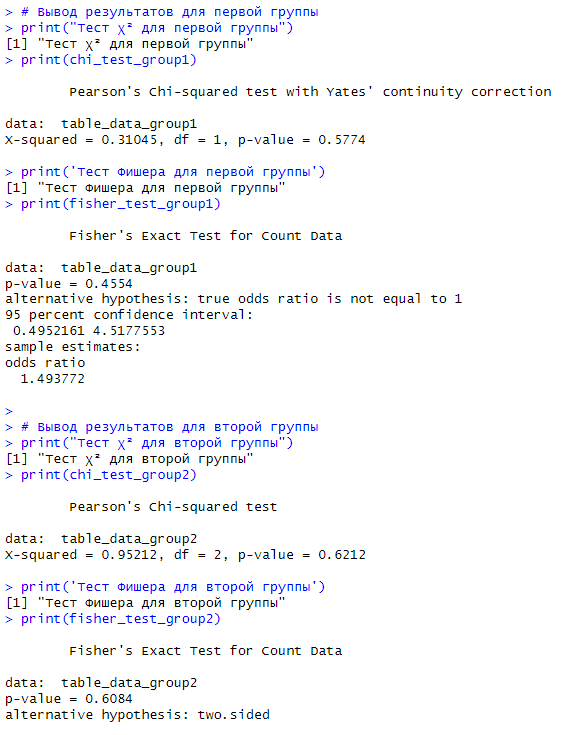


Рисунок 21. Результаты χ2 и Фишер

На рисунке 21 изображены данные после анализа χ2 и Фишера, в обеих группах показывается отсутствие статистически значимых связей между переменными. Это указывает на то, что переменные либо независимы, либо влияние одной переменной на другую слишком слабое, чтобы быть обнаруженным с помощью этих методов.

1. Дисперсионный анализа (ANOVA) и критерия Краскела-Уоллиса.

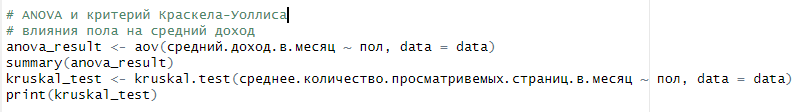


Рисунок 22. ANOVA и Краскел-Уоллис

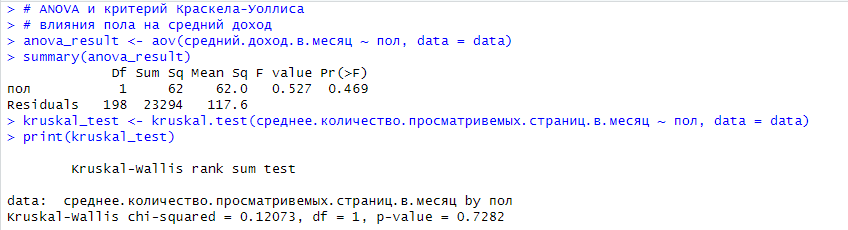


Рисунок 23. Результаты анализа

На рисунке 23 показаны результаты после анализа ANOVA и Краскел-Уоллис, где оба теста показывают, что пол не оказывает значимого влияния ни на средний доход, ни на среднее количество просматриваемых страниц в месяц. Это может означать, что в рамках данного набора данных мужчины и женщины имеют схожие средние значения как по доходам, так и по количеству посещаемых страниц, и различия между ними статистически незначимы.

1. Расчета коэффициентов корреляции Пирсона, Спирмена, Кендалла для первой и второй групп.

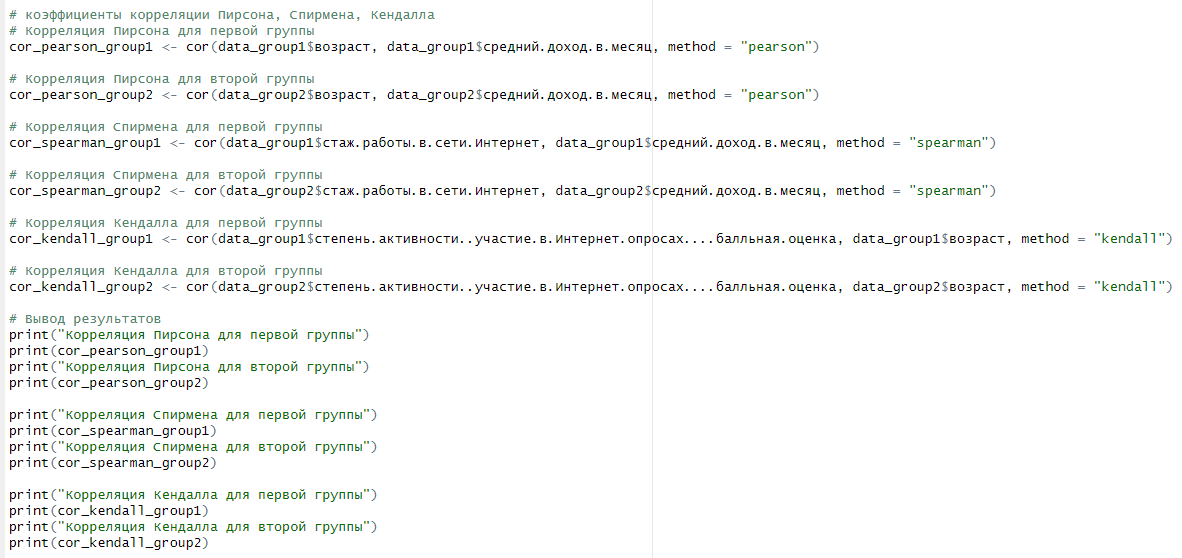


Рисунок 24. коэффициентов корреляции Пирсона, Спирмена, Кендалла

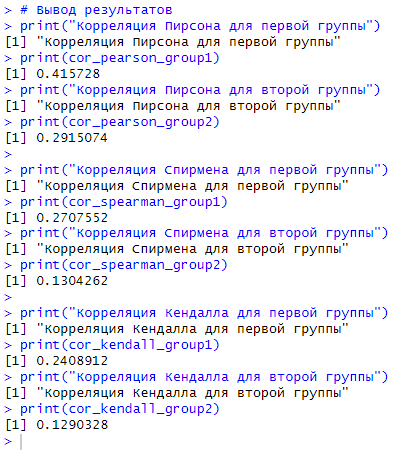


Рисунок 25. Результаты расчета

На рисунке 25 показаны результаты расчета коэффициентов корреляции Пирсона, Спирмена, Кендалла. Связь между возрастом и средним доходом существует, но она умеренная в первой группе и слабая во второй группе. Стаж работы в интернете оказывает слабое влияние на доход в обеих группах, а во второй группе эта связь еще слабее. Степень активности в интернет-опросах с возрастом слабо увеличивается, но эта связь едва заметна, особенно во второй группе. В целом, можно сказать, что возраст и стаж работы в интернете имеют некоторую, но не сильную связь с доходом и активностью в сети.

4. Расчет частного коэффициента корреляции для первой и второй группы



Рисунок 26. Частный коэффициент корреляции

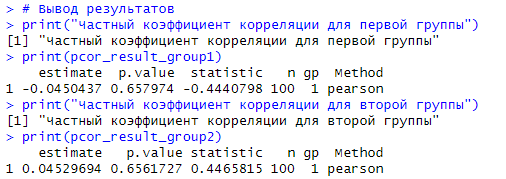


Рисунок 27. Результаты расчета

На рисунке 27 показаны результаты расчета частного коэффициента корреляции. В обеих группах наблюдаются слабые корреляционные связи (отрицательная в первой группе и положительная во второй), но они статистически незначимы, так как p-value значительно выше 0.05.

5. Графическое представление матрицы коэффициентов корреляции.

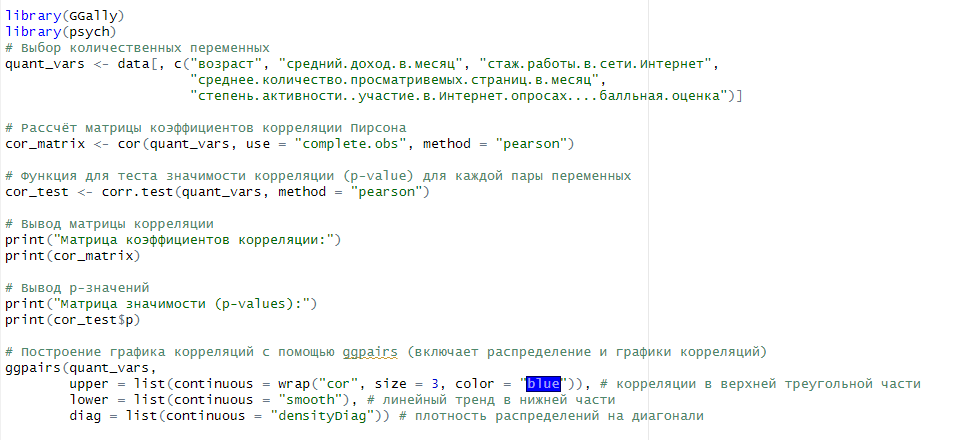


Рисунок 28. Графическое представление матрицы

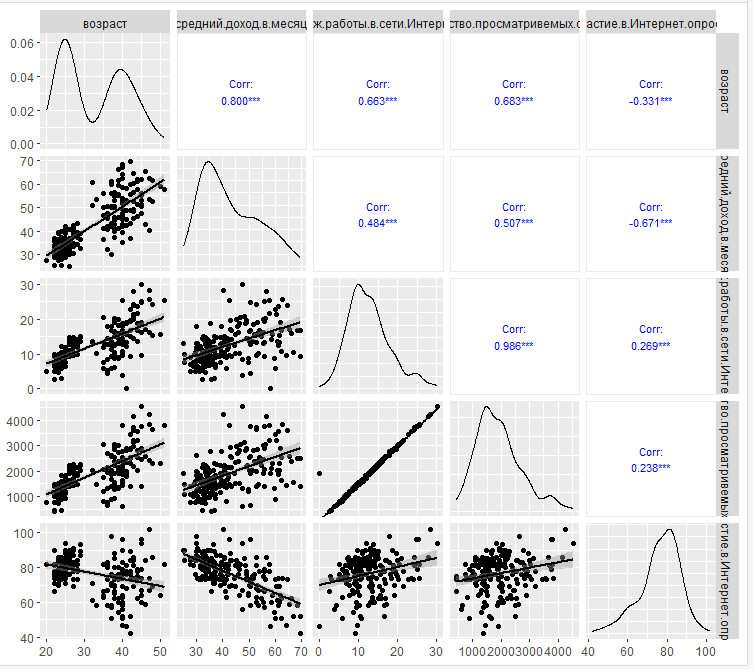


Рисунок 29. Графическое представление матрицы

На рисунке 29 изображена матрица коэффициентов корреляции, где возраст и доход имеют сильную положительную связь, стаж работы в интернете также коррелирует с доходом, но слабее, чем возраст. Количество просмотренных страниц сильно связано со стажем работы в интернете, что логично, так как с ростом опыта использования интернета пользователи больше времени проводят в сети. Участие в интернет-опросах снижено у людей с большим доходом и возрастом, что может указывать на снижение интереса к таким опросам с ростом дохода или занятости.

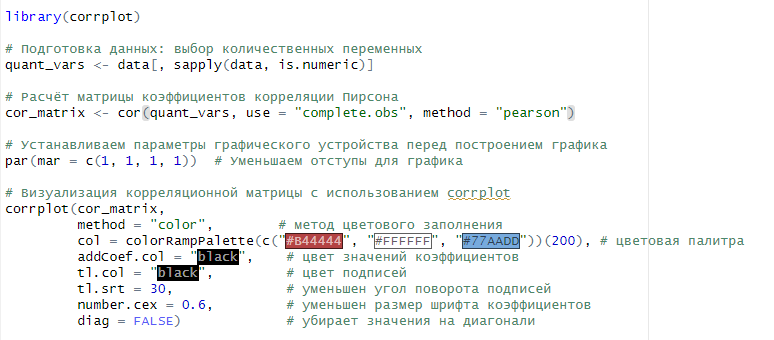


Рисунок 30. Графическое представление матрицы



Рисунок 31. Графическое представление матрицы

На рисунке 31 показана матрица коэффициентов корреляции Пирсона, в каждой ячейке диаграммы указано значение корреляции между двумя соответствующими характеристиками. Цвет ячейки отражает силу корреляции:

- Синий цвет - положительная корреляция (чем светлее, тем слабее).

- Красный цвет - отрицательная корреляция (чем темнее, тем слабее).

- Белый цвет - отсутствие корреляции.

**Заключение**

В ходе выполнения лабораторной работы были изучены базовые навыки работы в среде R, а также проведен анализ данных.

На основе данных сформирован общий вывод:

* Возраст, доход, стаж работы в Интернете и участие в Интернет-активностях тесно взаимосвязаны. Пользователи с более высоким доходом и стажем работы в сети более активно используют Интернет, но с возрастом и ростом дохода участие в интерактивных активностях, таких как опросы, снижается.
* Важность каждого фактора может варьироваться в зависимости от группы пользователей, что видно из частичных коэффициентов корреляции для разных групп, но общие тенденции сохраняются.

**Приложение**

#загрузка данных

data <- read.table(file = "datafile.csv", header = TRUE, sep = ";", fileEncoding = "Windows-1251",

na.strings = c("NA", "", " "))

# Просмотр первых строк таблицы

head(data)

# Просмотр таблицы

View(data)

# Имена столбцов

names(data)

# просмотр данных таблицы возраст

data$возраст

# подвыборка возраста

subset(data, data$возраст > 50)

# структура данных

str(data)

# Просмотр статистики по данным

summary(data)

# Минимальное значение возраста

min(data$возраст, na.rm = TRUE)

# Максимальное значение возраста

max(data$возраст, na.rm = TRUE)

# Среднее значение возраста

mean(data$возраст, na.rm = TRUE)

# Стандартное отклонение возраста

sd(data$возраст, na.rm = TRUE)

#Медиана

median(data$возраст, na.rm = TRUE)

# Квартильные значения (первая и третья квартили)

quantile(data$возраст, probs = c(0.25, 0.75), na.rm = TRUE)

# Мода

mode\_age <- as.numeric(names(sort(table(data$возраст), decreasing = TRUE)[1]))

print(mode\_age)

# Асимметрия packages("e1071")

library(e1071)

skewness(data$возраст, na.rm = TRUE)

# Эксцесс packages("e1071")

kurtosis(data$возраст, na.rm = TRUE)

# Диаграмма рассеяния возраста и дохода

plot(data$возраст, data$средний.доход.в.месяц,

xlab = "Возраст", ylab = "Средний доход в месяц",

main = "Диаграмма рассеяния: Возраст vs Средний доход")

# Радиальная диаграмма для качественного признака packages(plotrix)

survey\_table <- table(data$степень.активности..участие.в.Интернет.опросах....качественная.оценка, useNA = "no")

print(survey\_table)

library(plotrix)

survey\_values <- as.numeric(survey\_table)

survey\_labels <- names(survey\_table)

# Построение радиальной диаграммы

radial.plot(survey\_values,

labels = survey\_labels,

main = "Радиальная диаграмма по степени активности в опросах",

radial.lim = c(0, max(survey\_values)), # Установка пределов радиуса

rp.type = "p",

line.col = "blue", # Цвет линий

point.col = "red", # Цвет точек

show.grid.labels = TRUE) # Показывать подписи сетки

# Категориальная радиальная диаграмма по групповым переменным

library(plotrix)

# Создаем таблицу частот по профессиональной специализации и полу

category\_table <- table(data$профессиональная.специализация..насколько.тесно.профессия.клиента.связана.с.Интернет.,

data$пол)

category\_values\_male <- as.numeric(category\_table[, "1"])

category\_values\_female <- as.numeric(category\_table[, "2"])

# Определение меток для осей (профессиональная специализация)

labels <- rownames(category\_table)

# Построение радиальной диаграммы с учетом групповой переменной

radial.plot(category\_values\_male,

labels = labels,

main = "Категориальная радиальная диаграмма по профессиональной специализации и полу",

radial.lim = c(0, max(category\_values\_male, category\_values\_female)), # Установка пределов радиуса

rp.type = "p", # Тип графика "p" для точек

line.col = "blue",

point.col = "blue")

# Добавляем данные для женщин на тот же график

radial.plot(category\_values\_female,

add = TRUE, # Добавляем к существующему графику

line.col = "red",

point.col = "red")

legend("topright", legend = c("Мужской", "Женский"), col = c("blue", "red"), lty = 1, pch = 16)

# Категориальная столбиковая диаграмма для количественного признака в зависимости от групповой переменной

# Построение столбиковой диаграммы среднего дохода по полу

barplot(tapply(data$средний.доход.в.месяц, data$пол, mean, na.rm = TRUE),

main = "Средний доход по полу",

xlab = "Пол",

ylab = "Средний доход в месяц",

names.arg = c("Мужчины", "Женщины"), # Если 1 - мужчины, 2 - женщины

col = c("skyblue", "lightpink"),

ylim = c(0, max(tapply(data$средний.доход.в.месяц, data$пол, mean, na.rm = TRUE)) \* 1.1))

#Диаграмма размаха для количественного признака (возраст в зависимости от пола):

boxplot(возраст ~ пол, data = data,

main = "Диаграмма размаха возраста по полу",

xlab = "Пол",

ylab = "Возраст",

names = c("Мужчины", "Женщины"), # Если 1 - мужчины, 2 - женщины

col = c("skyblue", "lightpink")) # Цвет для каждого пола

#Гистограммы для количественных признаков:

# Настройка области графиков для отображения двух графиков в одном окне

par(mfrow = c(1, 2))

# Построение первой гистограммы для возраста

hist(data$возраст,

main = "Гистограмма возраста",

xlab = "Возраст",

col = "lightblue",

breaks = 10)

# Построение второй гистограммы для среднего дохода в месяц

hist(data$средний.доход.в.месяц,

main = "Гистограмма дохода",

xlab = "Средний доход в месяц",

col = "lightgreen",

breaks = 10)

# Возвращаем настройки к одному графику на окно

par(mfrow = c(1, 1))

# Матричный график для количественных переменных

pairs(~ возраст + средний.доход.в.месяц + стаж.работы.в.сети.Интернет,

data = data,

main = "Матричный график количественных переменных",

pch = 21,

bg = "lightblue",

col = "darkblue")

#Корреляционный анализ

# Фишера

# Подмножества для первой и второй группы

data\_group1 <- subset(data, группа == 1)

data\_group2 <- subset(data, группа == 2)

# Создание таблиц сопряжённости для первой группы

table\_data\_group1 <- table(data\_group1$пол, data\_group1$степень.активности..участие.в.Интернет.опросах....качественная.оценка)

chi\_test\_group1 <- chisq.test(table\_data\_group1)

fisher\_test\_group1 <- fisher.test(table\_data\_group1)

# Создание таблиц сопряжённости для второй группы

table\_data\_group2 <- table(data\_group2$пол, data\_group2$степень.активности..участие.в.Интернет.опросах....качественная.оценка)

chi\_test\_group2 <- chisq.test(table\_data\_group2)

fisher\_test\_group2 <- fisher.test(table\_data\_group2)

# Вывод результатов для первой группы

print("Тест χ² для первой группы")

print(chi\_test\_group1)

print('Тест Фишера для первой группы')

print(fisher\_test\_group1)

# Вывод результатов для второй группы

print("Тест χ² для второй группы")

print(chi\_test\_group2)

print('Тест Фишера для второй группы')

print(fisher\_test\_group2)

# ANOVA и критерий Краскела-Уоллиса

# влияния пола на средний доход

anova\_result <- aov(средний.доход.в.месяц ~ пол, data = data)

summary(anova\_result)

kruskal\_test <- kruskal.test(среднее.количество.просматривемых.страниц.в.месяц ~ пол, data = data)

print(kruskal\_test)

# коэффициенты корреляции Пирсона, Спирмена, Кендалла

# Корреляция Пирсона для первой группы

cor\_pearson\_group1 <- cor(data\_group1$возраст, data\_group1$средний.доход.в.месяц, method = "pearson")

# Корреляция Пирсона для второй группы

cor\_pearson\_group2 <- cor(data\_group2$возраст, data\_group2$средний.доход.в.месяц, method = "pearson")

# Корреляция Спирмена для первой группы

cor\_spearman\_group1 <- cor(data\_group1$стаж.работы.в.сети.Интернет, data\_group1$средний.доход.в.месяц, method = "spearman")

# Корреляция Спирмена для второй группы

cor\_spearman\_group2 <- cor(data\_group2$стаж.работы.в.сети.Интернет, data\_group2$средний.доход.в.месяц, method = "spearman")

# Корреляция Кендалла для первой группы

cor\_kendall\_group1 <- cor(data\_group1$степень.активности..участие.в.Интернет.опросах....балльная.оценка, data\_group1$возраст, method = "kendall")

# Корреляция Кендалла для второй группы

cor\_kendall\_group2 <- cor(data\_group2$степень.активности..участие.в.Интернет.опросах....балльная.оценка, data\_group2$возраст, method = "kendall")

# Вывод результатов

print("Корреляция Пирсона для первой группы")

print(cor\_pearson\_group1)

print("Корреляция Пирсона для второй группы")

print(cor\_pearson\_group2)

print("Корреляция Спирмена для первой группы")

print(cor\_spearman\_group1)

print("Корреляция Спирмена для второй группы")

print(cor\_spearman\_group2)

print("Корреляция Кендалла для первой группы")

print(cor\_kendall\_group1)

print("Корреляция Кендалла для второй группы")

print(cor\_kendall\_group2)

#частный коэффициент корреляции

# Подмножества для первой и второй группы

data\_group1 <- subset(data, группа == 1)

data\_group2 <- subset(data, группа == 2)

library(ppcor)

# Рассчитываем матрицу корреляции Пирсона для первой группы

cor\_matrix\_group1 <- cor(data\_group1[, c("возраст", "средний.доход.в.месяц", "стаж.работы.в.сети.Интернет")], use = "complete.obs")

# Находим пару переменных с максимальным модулем коэффициента корреляции

max\_corr\_vars\_group1 <- which(abs(cor\_matrix\_group1) == max(abs(cor\_matrix\_group1[upper.tri(cor\_matrix\_group1)])), arr.ind = TRUE)

# Повторяем для второй группы

cor\_matrix\_group2 <- cor(data\_group2[, c("возраст", "средний.доход.в.месяц", "стаж.работы.в.сети.Интернет")], use = "complete.obs")

max\_corr\_vars\_group2 <- which(abs(cor\_matrix\_group2) == max(abs(cor\_matrix\_group2[upper.tri(cor\_matrix\_group2)])), arr.ind = TRUE)

# Частный коэффициент корреляции для первой группы

var1\_group1 <- names(data\_group1)[max\_corr\_vars\_group1[1]]

var2\_group1 <- names(data\_group1)[max\_corr\_vars\_group1[2]]

pcor\_result\_group1 <- pcor.test(data\_group1[[var1\_group1]], data\_group1[[var2\_group1]], data\_group1$стаж.работы.в.сети.Интернет)

# Частный коэффициент корреляции для второй группы

var1\_group2 <- names(data\_group2)[max\_corr\_vars\_group2[1]]

var2\_group2 <- names(data\_group2)[max\_corr\_vars\_group2[2]]

pcor\_result\_group2 <- pcor.test(data\_group2[[var1\_group2]], data\_group2[[var2\_group2]], data\_group2$стаж.работы.в.сети.Интернет)

# Вывод результатов

print("Частный коэффициент корреляции для первой группы")

print(pcor\_result\_group1)

print("Частный коэффициент корреляции для второй группы")

print(pcor\_result\_group2)

library(GGally)

library(psych)

# Выбор количественных переменных

quant\_vars <- data[, c("возраст", "средний.доход.в.месяц", "стаж.работы.в.сети.Интернет",

"среднее.количество.просматривемых.страниц.в.месяц",

"степень.активности..участие.в.Интернет.опросах....балльная.оценка")]

# Рассчёт матрицы коэффициентов корреляции Пирсона

cor\_matrix <- cor(quant\_vars, use = "complete.obs", method = "pearson")

# Функция для теста значимости корреляции (p-value) для каждой пары переменных

cor\_test <- corr.test(quant\_vars, method = "pearson")

# Вывод матрицы корреляции

print("Матрица коэффициентов корреляции:")

print(cor\_matrix)

# Вывод p-значений

print("Матрица значимости (p-values):")

print(cor\_test$p)

# Построение графика корреляций с помощью ggpairs (включает распределение и графики корреляций)

ggpairs(quant\_vars,

upper = list(continuous = wrap("cor", size = 3, color = "blue")), # корреляции в верхней треугольной части

lower = list(continuous = "smooth"), # линейный тренд в нижней части

diag = list(continuous = "densityDiag")) # плотность распределений на диагонали

library(corrplot)

# Подготовка данных: выбор количественных переменных

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

# Расчёт матрицы коэффициентов корреляции Пирсона

cor\_matrix <- cor(quant\_vars, use = "complete.obs", method = "pearson")

# Устанавливаем параметры графического устройства перед построением графика

par(mar = c(1, 1, 1, 1)) # Уменьшаем отступы для графика

# Визуализация корреляционной матрицы с использованием corrplot

corrplot(cor\_matrix,

method = "color", # метод цветового заполнения

col = colorRampPalette(c("#B44444", "#FFFFFF", "#77AADD"))(200), # цветовая палитра

addCoef.col = "black", # цвет значений коэффициентов

tl.col = "black", # цвет подписей

tl.srt = 30, # уменьшен угол поворота подписей

number.cex = 0.6, # уменьшен размер шрифта коэффициентов

diag = FALSE) # убирает значения на диагонали