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 32](#_bookmark2)

[Приложение 33](#_bookmark3)

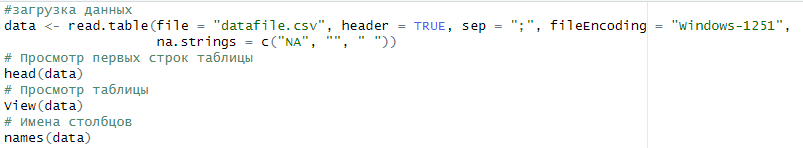
# **Постановка задачи**

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.

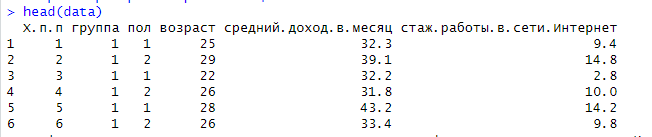


Рисунок 2. Результат работы head

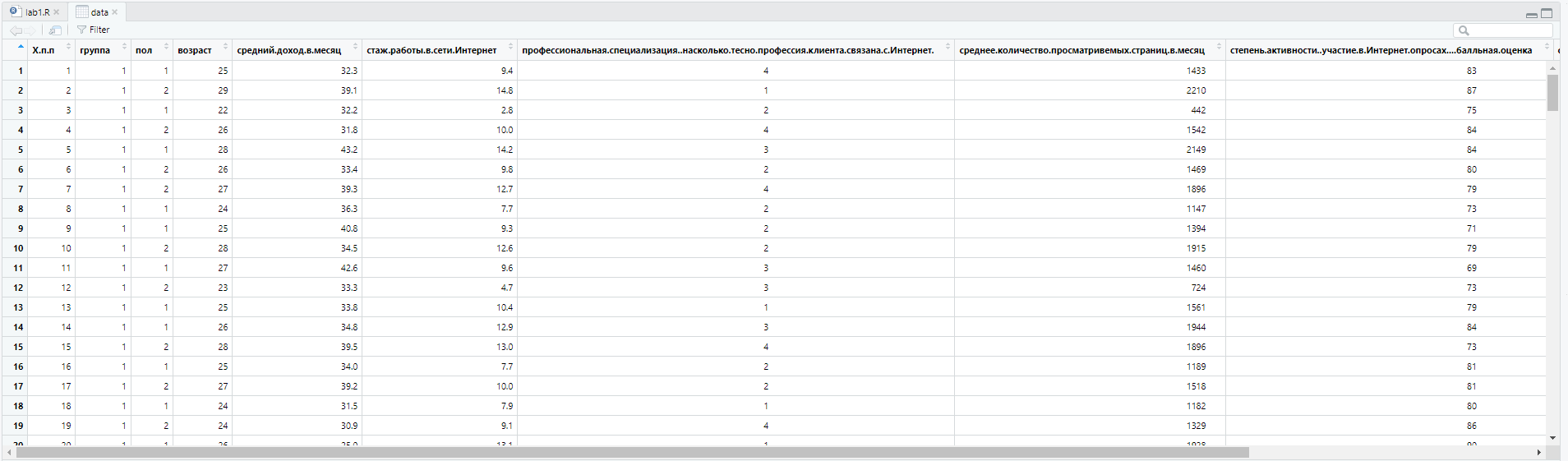


Рисунок 3. Результат работы view

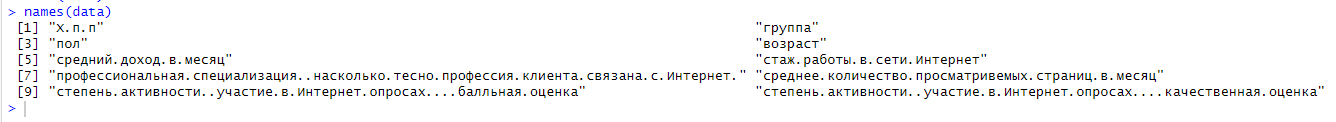


Рисунок 4. Результат работы names

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

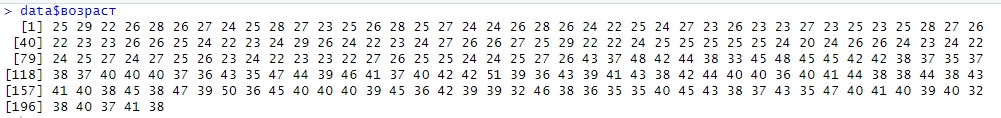


Рисунок 5. Результат работы data$возраст

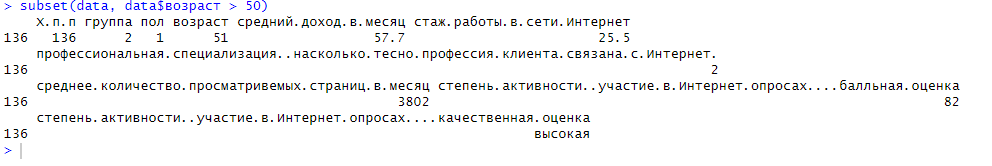


Рисунок 6. Результат работы subset(data, data$возраст > 50)

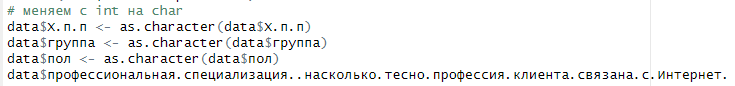


Рисунок 7. Смена типа данных

Смена типа данных необходима для корректного анализа.

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 8.

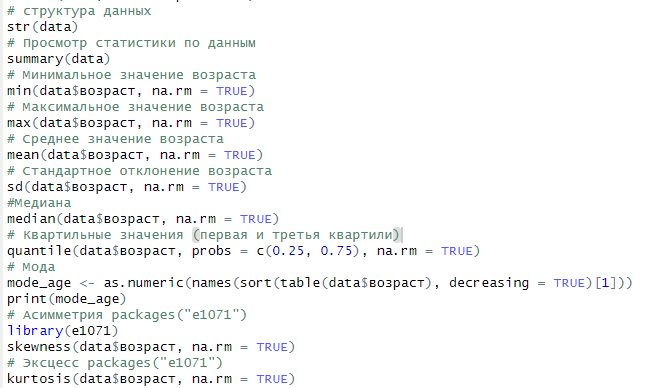


Рисунок 8. Команды для описания основных статистических характеристик

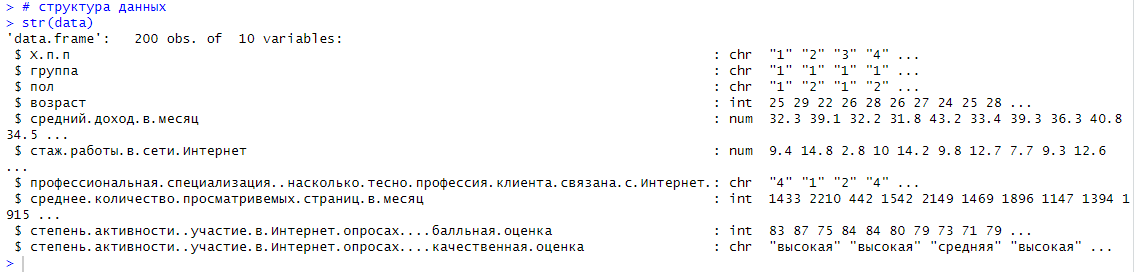


Рисунок 9. Структура данных

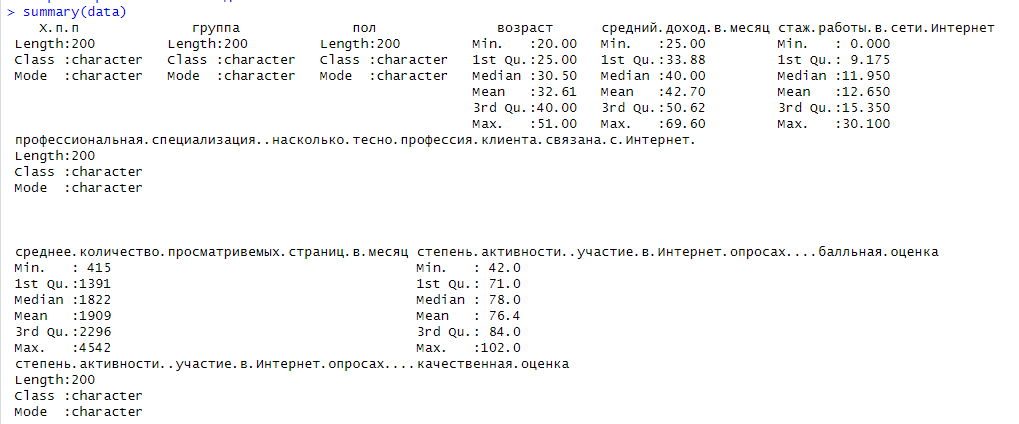


Рисунок 10. Статистика данных

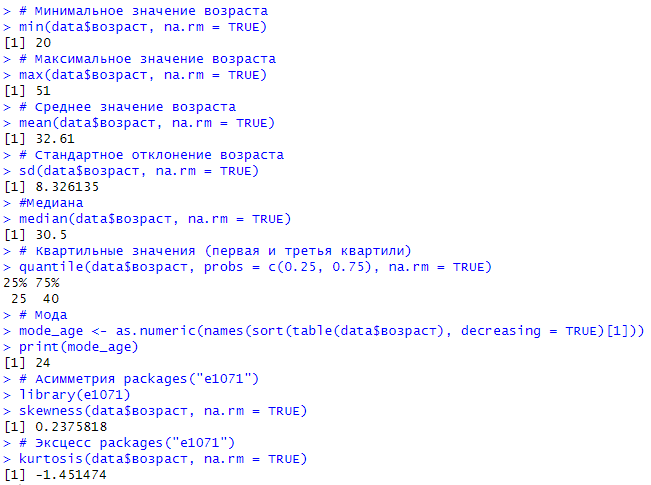


Рисунок 11. Значения по возрасту

На рисунке 11 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 12 – 15 показан код для построения диаграмм.



Рисунок 12. Описание диаграмм

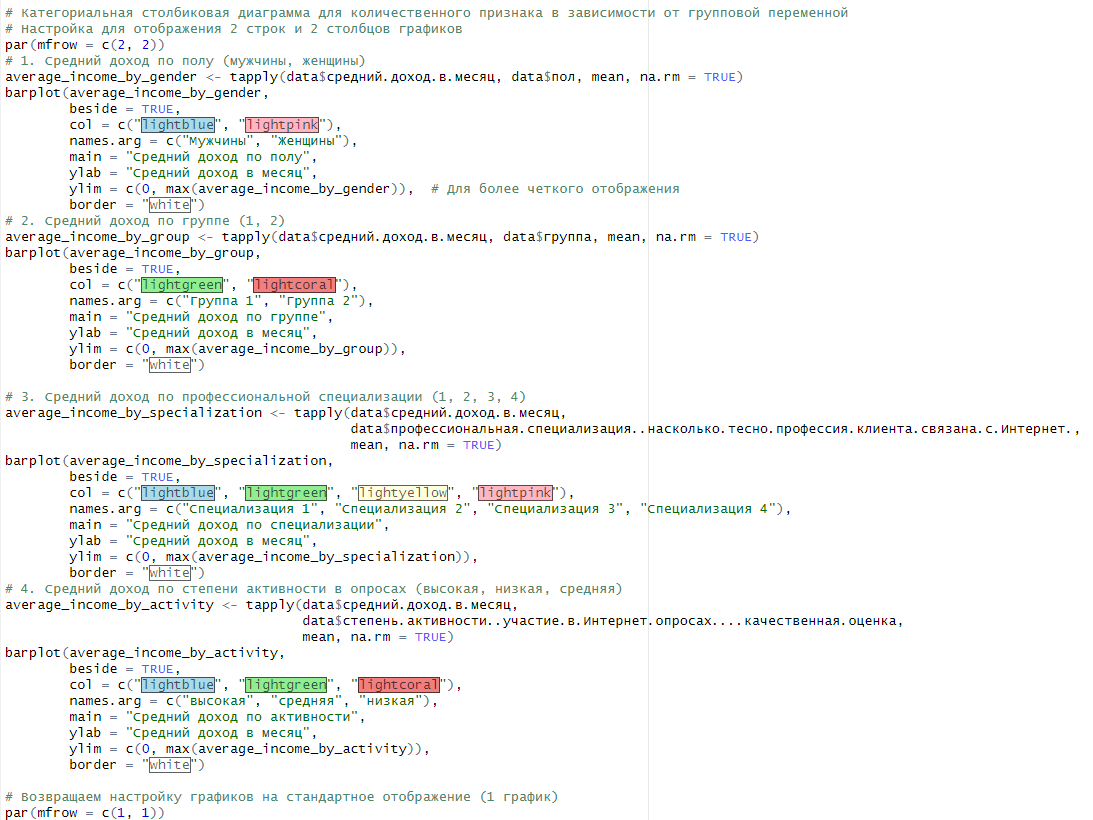


Рисунок 13. Описание диаграмм

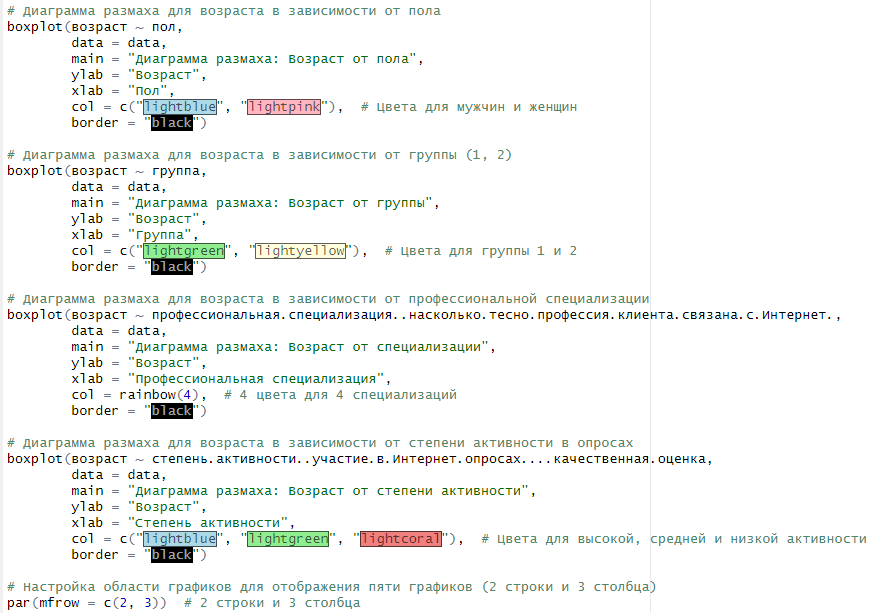


Рисунок 14. Описание диаграмм



Рисунок 15. Описание диаграмм

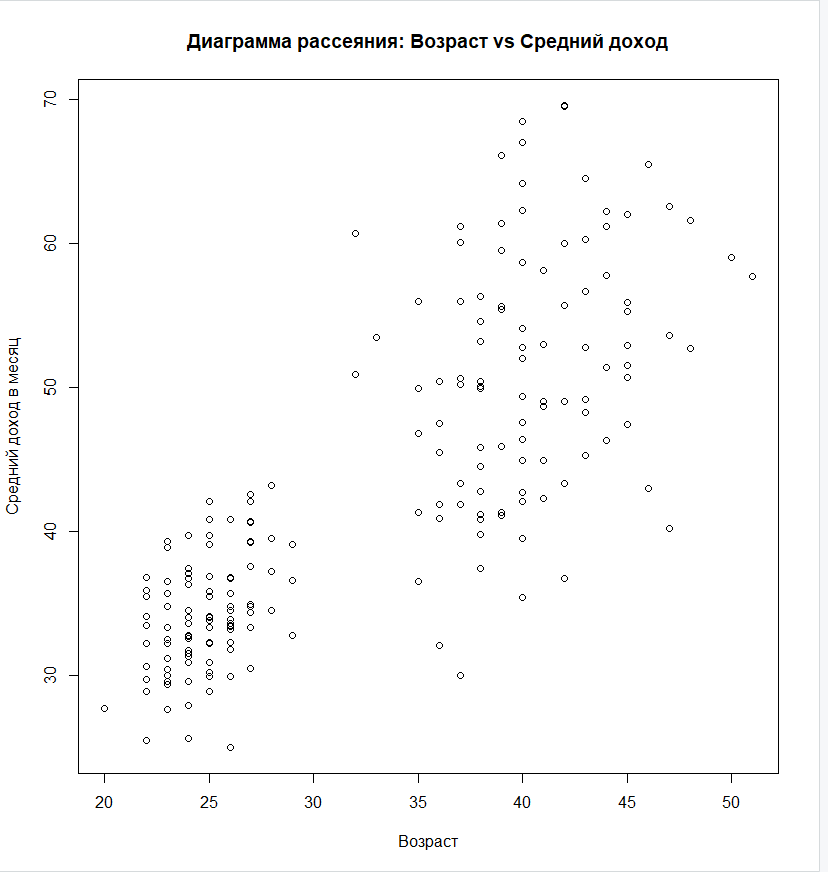


Рисунок 16. Диаграмма рассеяния

На рисунке 16 показана диаграмма рассеяния возраста и среднего дохода в месяц. Диаграмма показывает, что средний доход в целом увеличивается с возрастом, хотя есть значительные различия между людьми в любом данном возрасте, особенно в основные рабочие годы - в 30-е и 40-е годы.

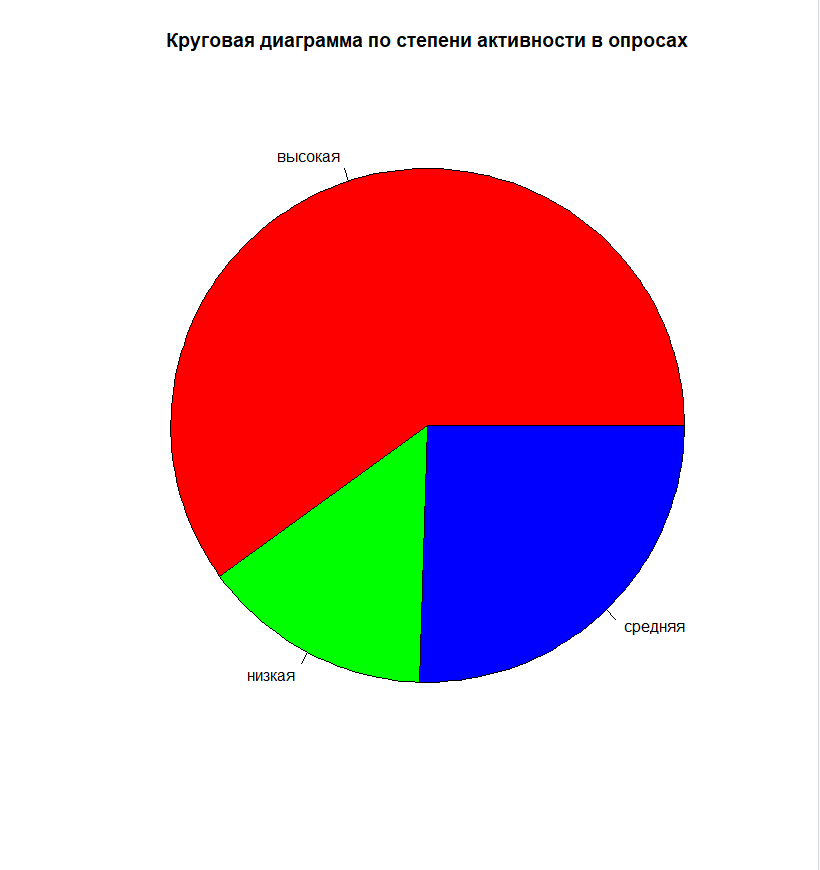


Рисунок 17. Радиальная диаграмма

На рисунке 17 изображена радиальная диаграмма, показывающая степень активности участия в интернет-опросах.



Рисунок 18. Категориальная радиальная диаграмма

На рисунке 18 изображена категориальная радиальная диаграмма на основе пола людей, данная диаграмма показывает, как пол влияет на различные факторы.

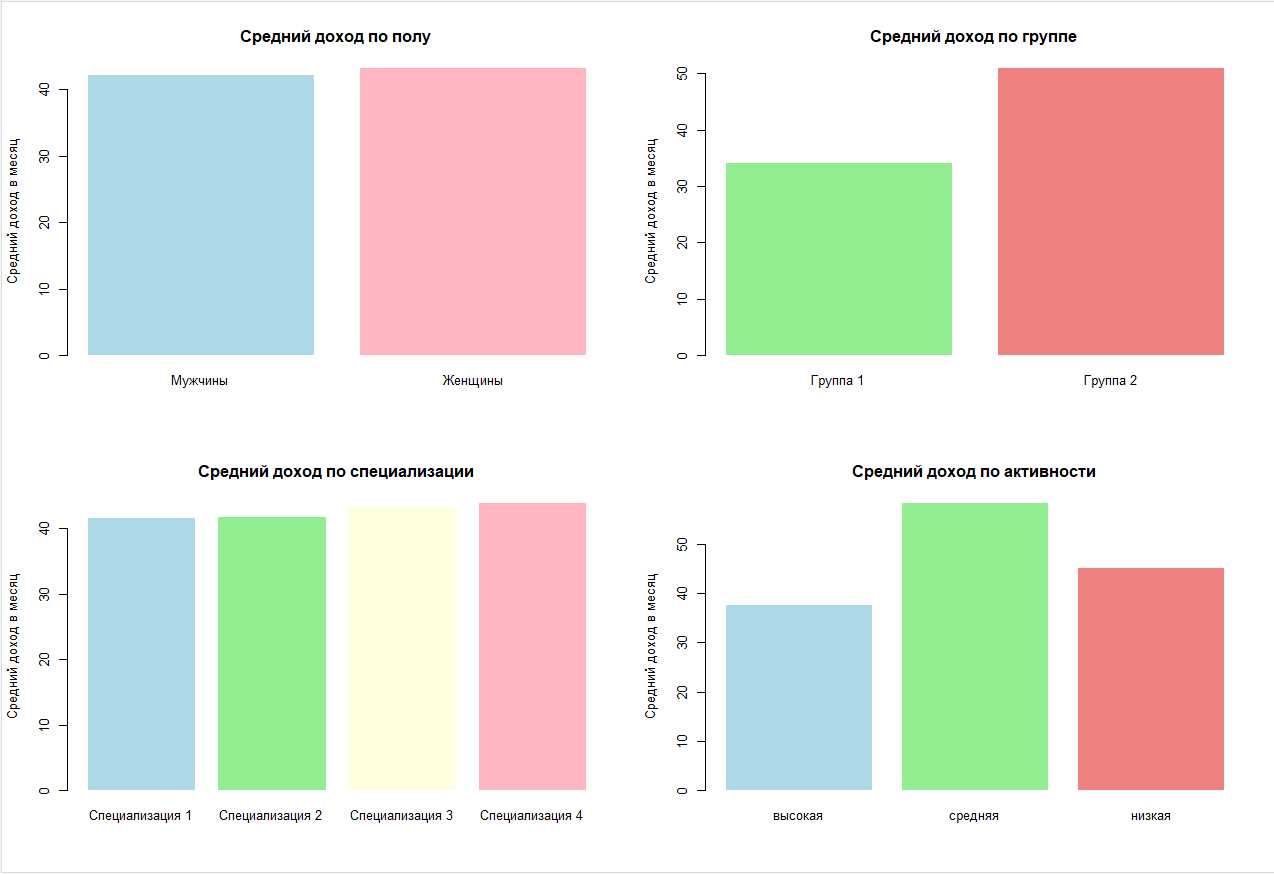


Рисунок 19. Категориальная столбиковая диаграмма

На рисунке 19 изображена категориальная столбиковая диаграмма, показывающая средний доход в зависимости от пола, группы, специализации, активности в опросах.

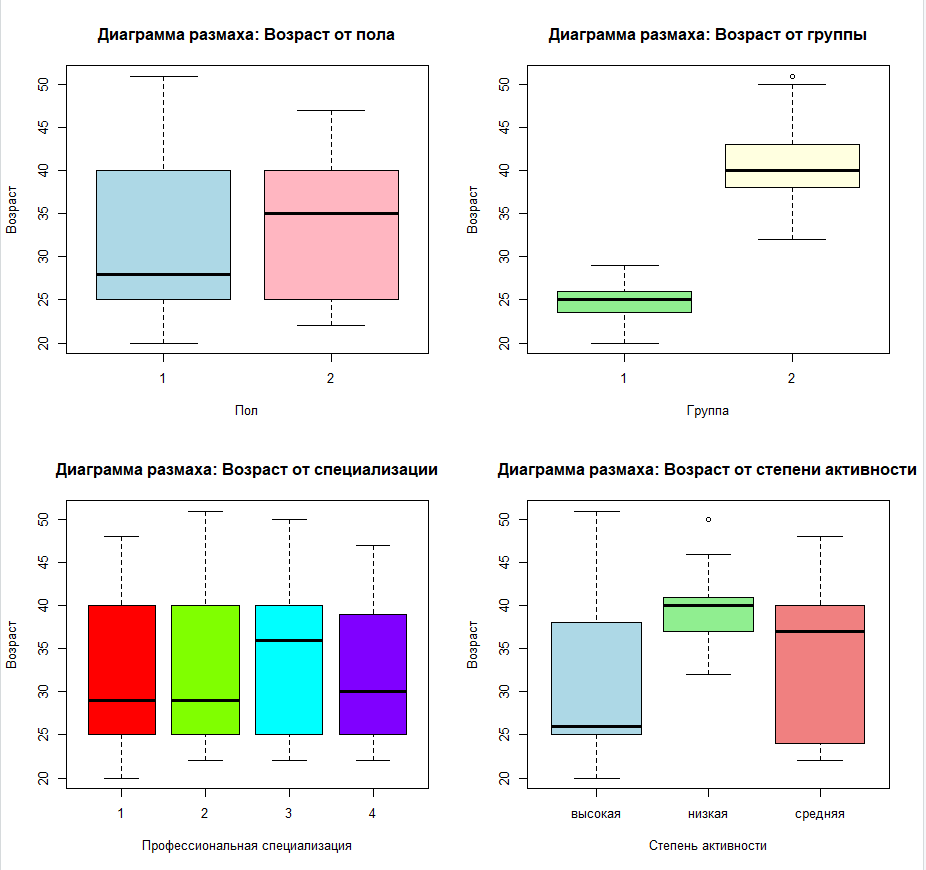


Рисунок 20. Диаграмма размаха

На рисунке 20 показаны диаграммы размаха возраст от пола, возраст от группы, возраст от специализации, возраст от степени активности в интернет-опросах.

* Возраст от пола: Средний возраст не сильно отличается между мужским и женским полом, мужской имеет более широкий диапазон возрастов.
* Возраст от группы: В группе 1 более молодые участники, чем в группе 2.
* Возраст от специализации: Средний возраст не сильно зависит от профессиональной специализации, поскольку медианы в четырех категориях близки друг к другу. Все четыре категории имеют примерно одинаковый возрастной разброс.
* Возраст от степени активности в интернет-опросах: Люди с высокой степенью активности имеют самый широкий возрастной разброс, и их возраст варьируется от самых молодых до самых старших значений. Медианный возраст участников с низкой активностью выше, чем у других категорий, что может свидетельствовать о том, что люди с низкой активностью, в среднем, старше.

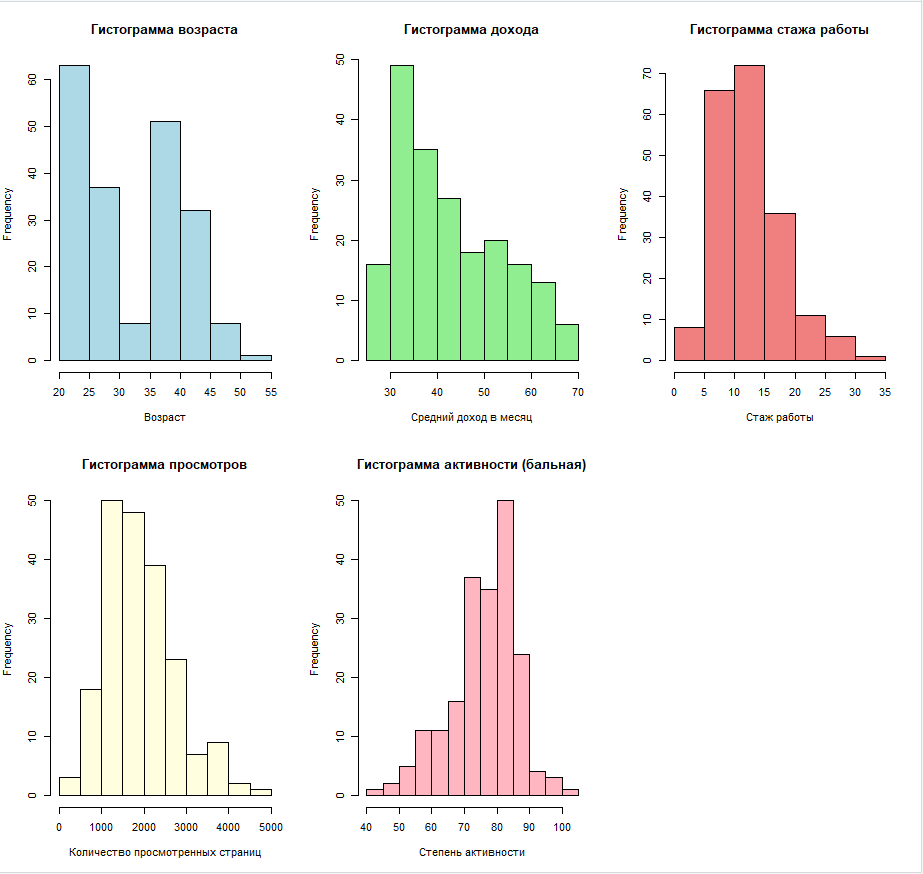


Рисунок 21. Гистограммы для количественных признаков

На рисунке 21 показаны гистограммы:

* Гистограмма возраста показывает, что большинство людей в выборке находятся в возрасте от 25 до 45 лет, с небольшим пиком около 30 лет.
* Гистограмма дохода демонстрирует, что наибольшее количество людей имеет средний доход в месяц в диапазоне от 30 до 50 условных единиц.
* Гистограмма стажа работы отражает, что значительная часть людей имеет относительно небольшой стаж - от 0 до 10 лет. Стаж более 20–25 лет встречается гораздо реже.
* Количество просмотренных страниц большинством пользователей находится в диапазоне от 1000 до 2000. Просмотры свыше 3000 страниц совершает лишь малая часть людей.
* Гистограмма активности показывает, что степень активности большинства пользователей варьируется от 50 до 80 баллов из 100.

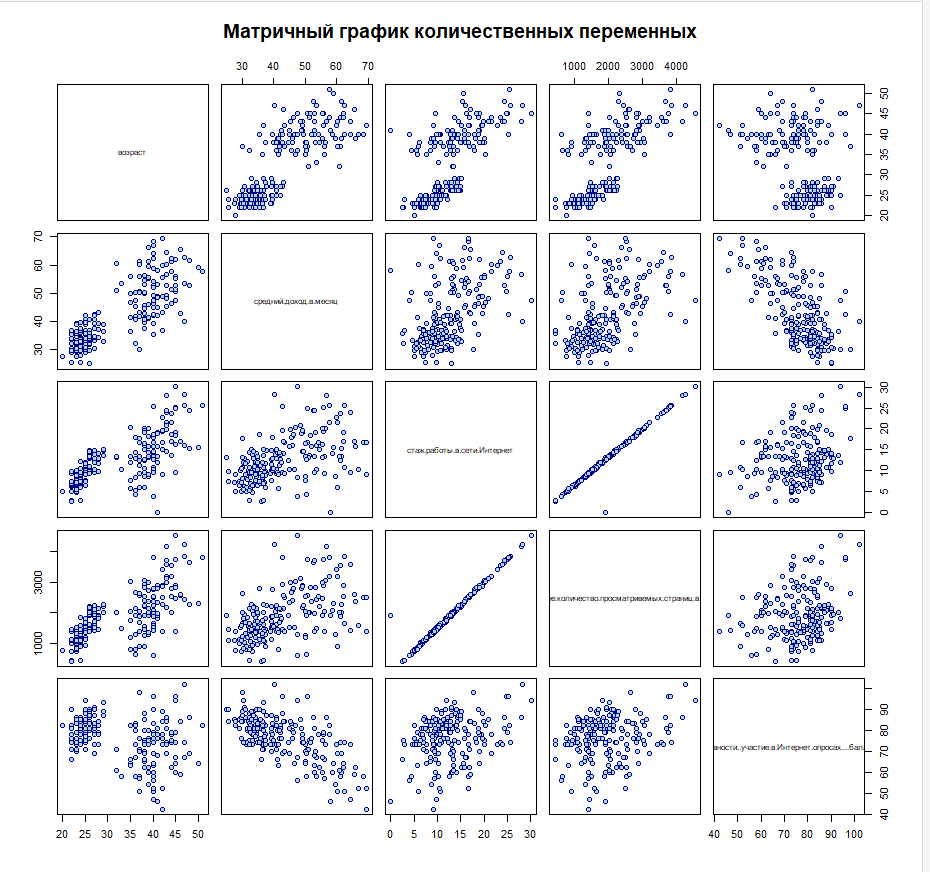


Рисунок 22. Матричный график для количественных переменных

На графике показывается диаграммы рассеивания между количественными переменами.

Далее был проведен корреляционный анализ.

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

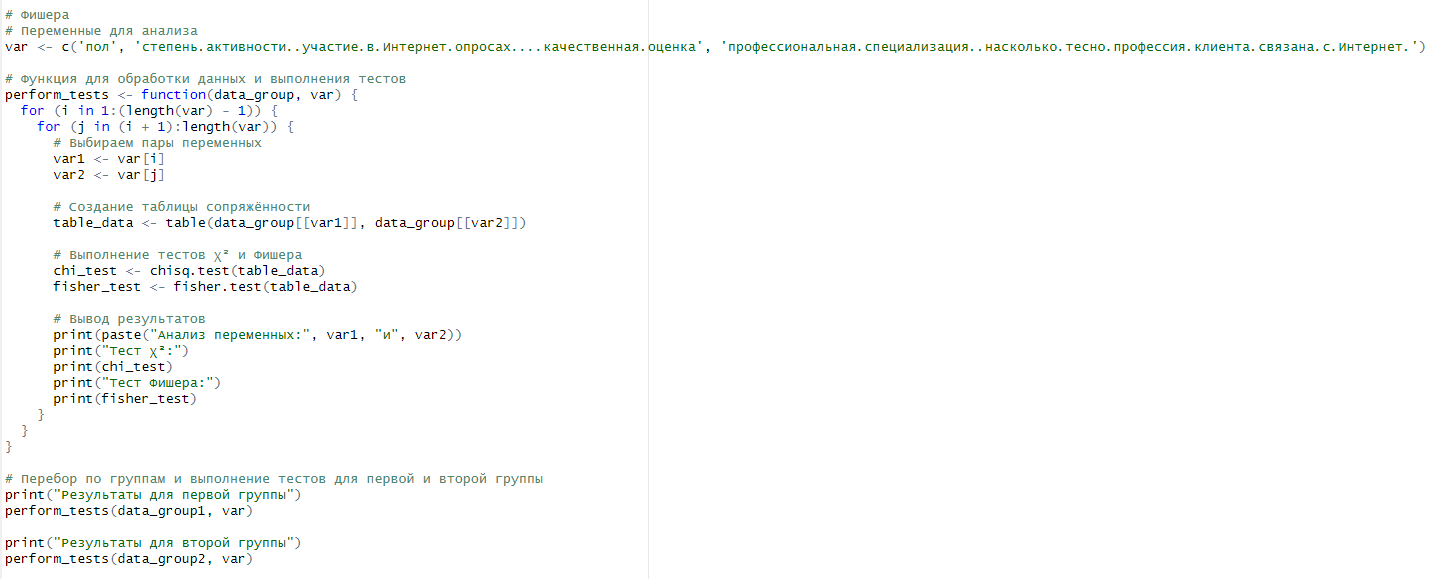
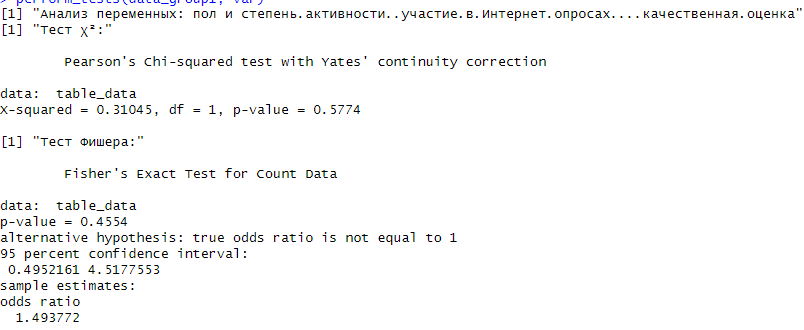
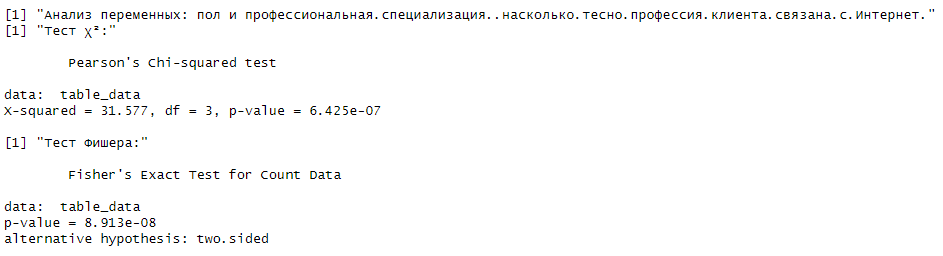
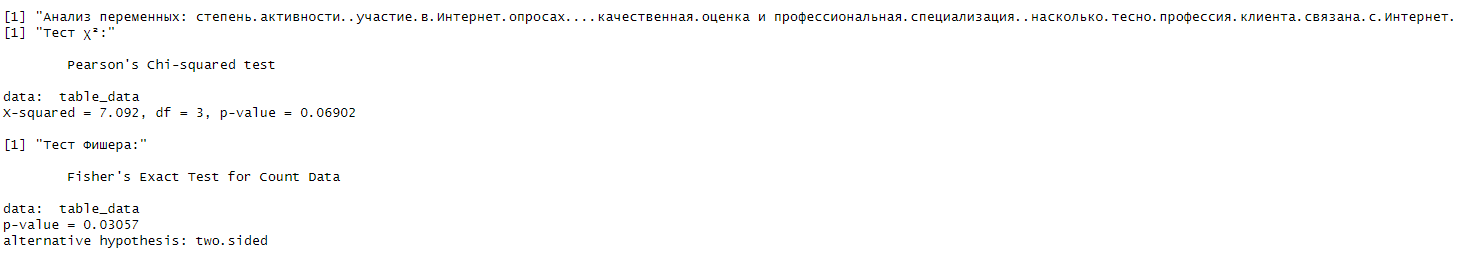


Рисунок 23. χ2 и Фишер

Результаты первой группы:







Результаты анализа взаимосвязи между переменными "пол" и "степень активности участие в Интернет-опросах, качественная оценка":

* Тест χ² (хи-квадрат): p-значение равно 0.5774, что больше общепринятого уровня значимости 0.05. Это означает, что нет статистически значимой взаимосвязи между переменными "пол" и "степень активности" на основе теста χ².
* Результаты точного теста Фишера также указывают на отсутствие статистически значимой взаимосвязи между переменными, так как p-значение равно 0.4554, что также больше уровня значимости 0.05.

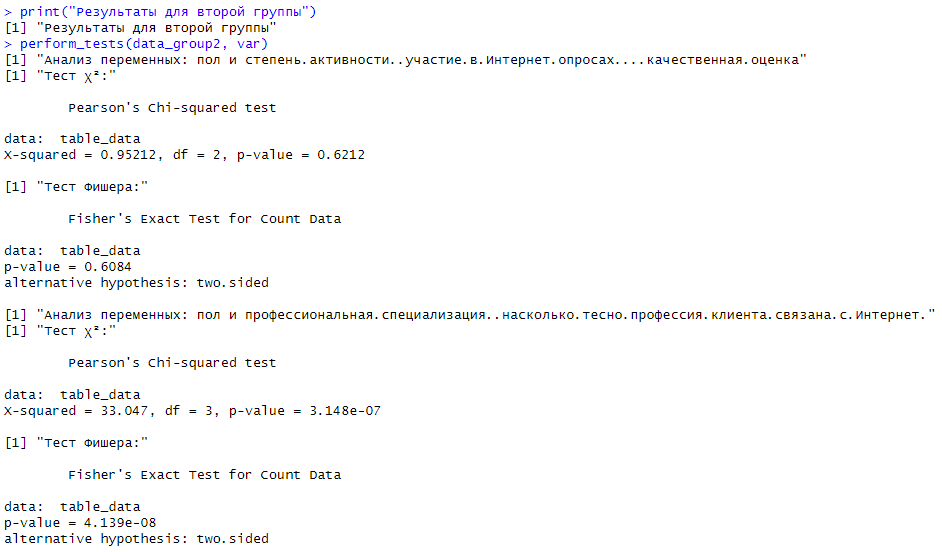
Результаты анализа взаимосвязи между переменными "пол" и " профессиональная специализация":

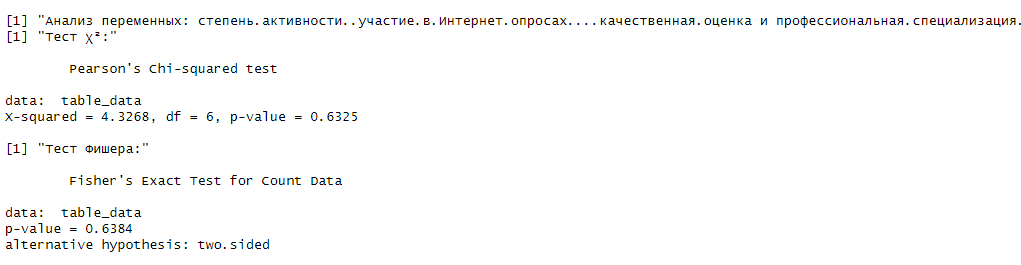
* Тест χ² (хи-квадрат): p-значение равно 6.425e-07, что значительно меньше уровня значимости 0.05. Это указывает на наличие статистически значимой взаимосвязи.
* Результаты точного теста Фишера подтверждают выводы теста χ² о наличии статистически значимой взаимосвязи, так как p-значение равно 8.913e-08, что также значительно меньше уровня значимости 0.05

Результаты анализа взаимосвязи между переменными " степень активности в участии в интернет-опросах" и "профессиональная специализация":

* Тест χ² (хи-квадрат): p-значение равно 0.06902, что превышает уровень значимости 0.05. Это означает, что нет статистически значимой взаимосвязи.
* Результаты теста Фишера указывают на наличие статистически значимой взаимосвязи, так как p-значение равно 0.03057, что меньше уровня значимости 0.05.
* Результаты теста χ² не подтверждают наличие значимой взаимосвязи между степенью активности в интернет-опросах и профессиональной специализацией. Однако результаты точного теста Фишера показывают, что такая взаимосвязь существует. Это может указывать на то, что в некоторых случаях (например, в определённых профессиях) активность в интернет-опросах действительно может зависеть от профессиональной специализации.

Результаты второй группы:





На основе результатов первой и второй группы можно сделать вывод, что:

1. Связь между полом и степенью активности: В обеих группах отсутствует статистически значимая связь.

2. Связь между полом и профессиональной специализацией: В обеих группах имеется статистически значимая связь.

3. Связь между степенью активности и профессиональной специализацией: В первой группе есть признаки связи (по тесту Фишера), а во второй группе связь отсутствует.

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

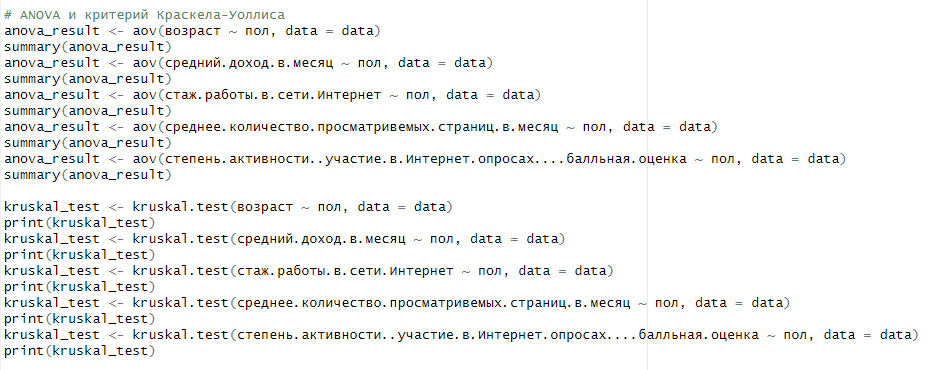
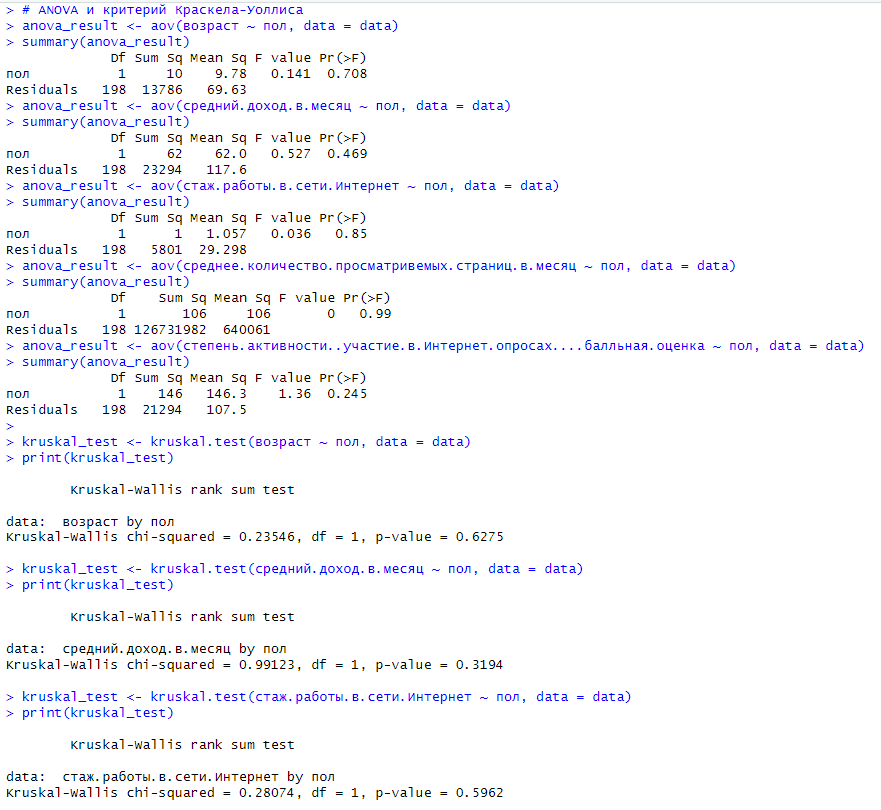


Рисунок 24. ANOVA и Краскел-Уоллис



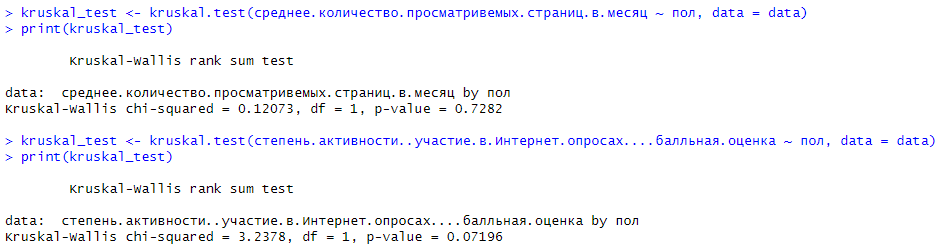


Рисунок 25. Результаты анализа

На рисунке 25 показаны результаты методов ANOVA и Краскел-Уоллиса о влиянии пола на различные характеристики:

* Влияние на возраст: ANOVA: p = 0.708, Краскал-Уоллис: p = 0.6275. В обоих тестах не обнаружено статистически значимых различий в возрасте между полами (p> 0.05).
* Влияние на средний доход: ANOVA: p = 0.469, Краскал-Уоллис: p = 0.3194. Нет статистически значимых различий в среднем доходе в месяц между полами (p > 0.05) по обоим тестам.
* Влияние на стаж работы: ANOVA: p = 0.85, Краскал-Уоллис: p = 0.5962. Оба теста показывают отсутствие статистически значимых различий в стаже работы в сети Интернет между полами (p > 0.05).
* Влияние на среднее количество просматриваемых страниц в месяц: ANOVA: p = 0.99, Краскал-Уоллис: p = 0.7282. Нет статистически значимых различий в среднем количестве просматриваемых страниц в месяц между полами (p > 0.05).
* Влияние на участие в интернет-опросах: ANOVA: p = 0.245, Краскал-Уоллис: p = 0.07196. ANOVA не показывает значимости (p > 0.05), однако тест Краскала-Уоллиса близок к значимости (p ≈ 0.072). Это может указывать на потенциальное различие, но оно не является статистически значимым на уровне α = 0.05.

Таким образом, можно сделать вывод, что пол не оказывает значительного влияния на рассматриваемые переменные.

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



Рисунок 26. Расчет коэффициентов корреляции Пирсона, Спирмена, Кендалла

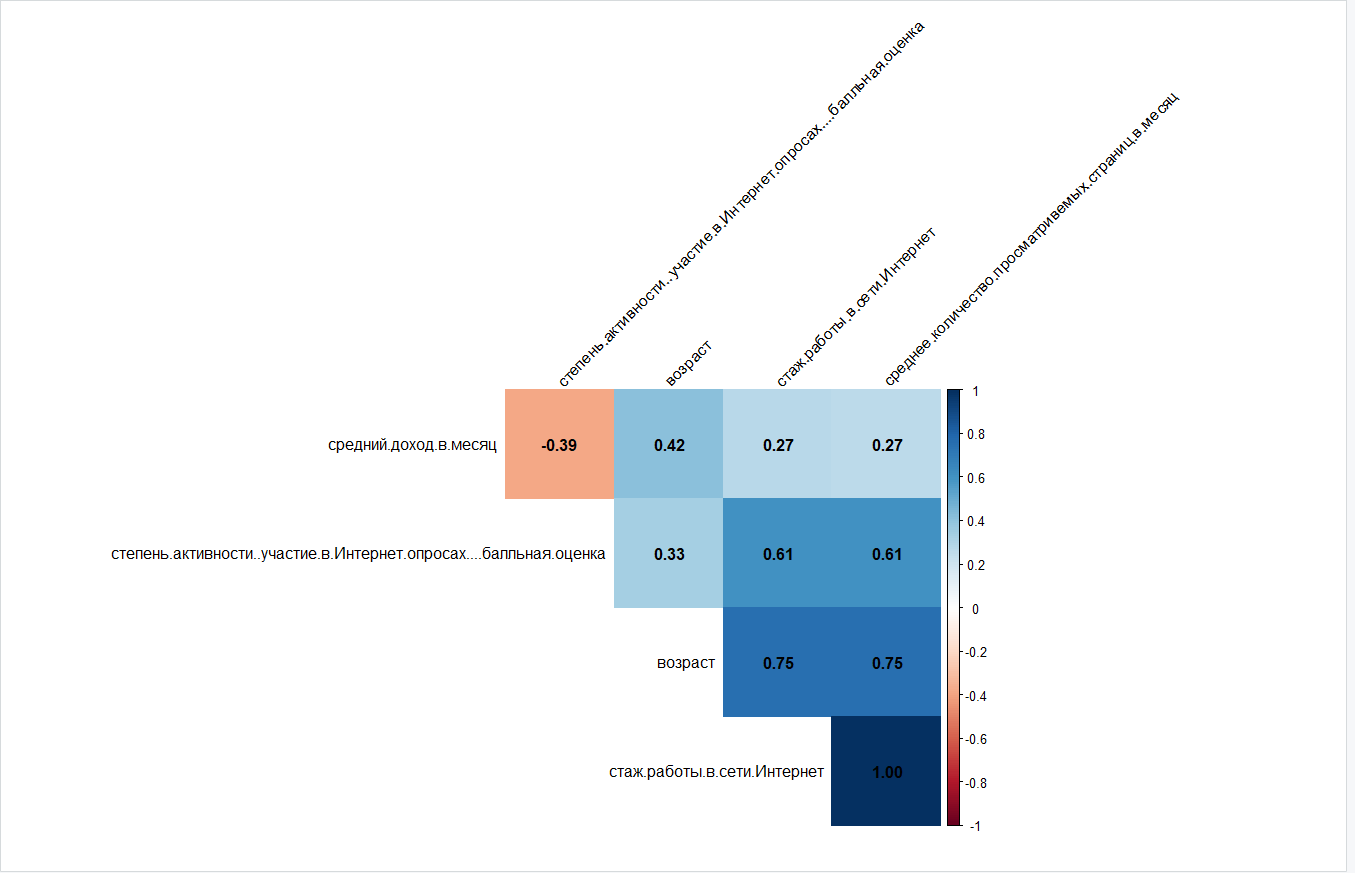


Рисунок 27. Корреляция методом Пирсона для первой группы

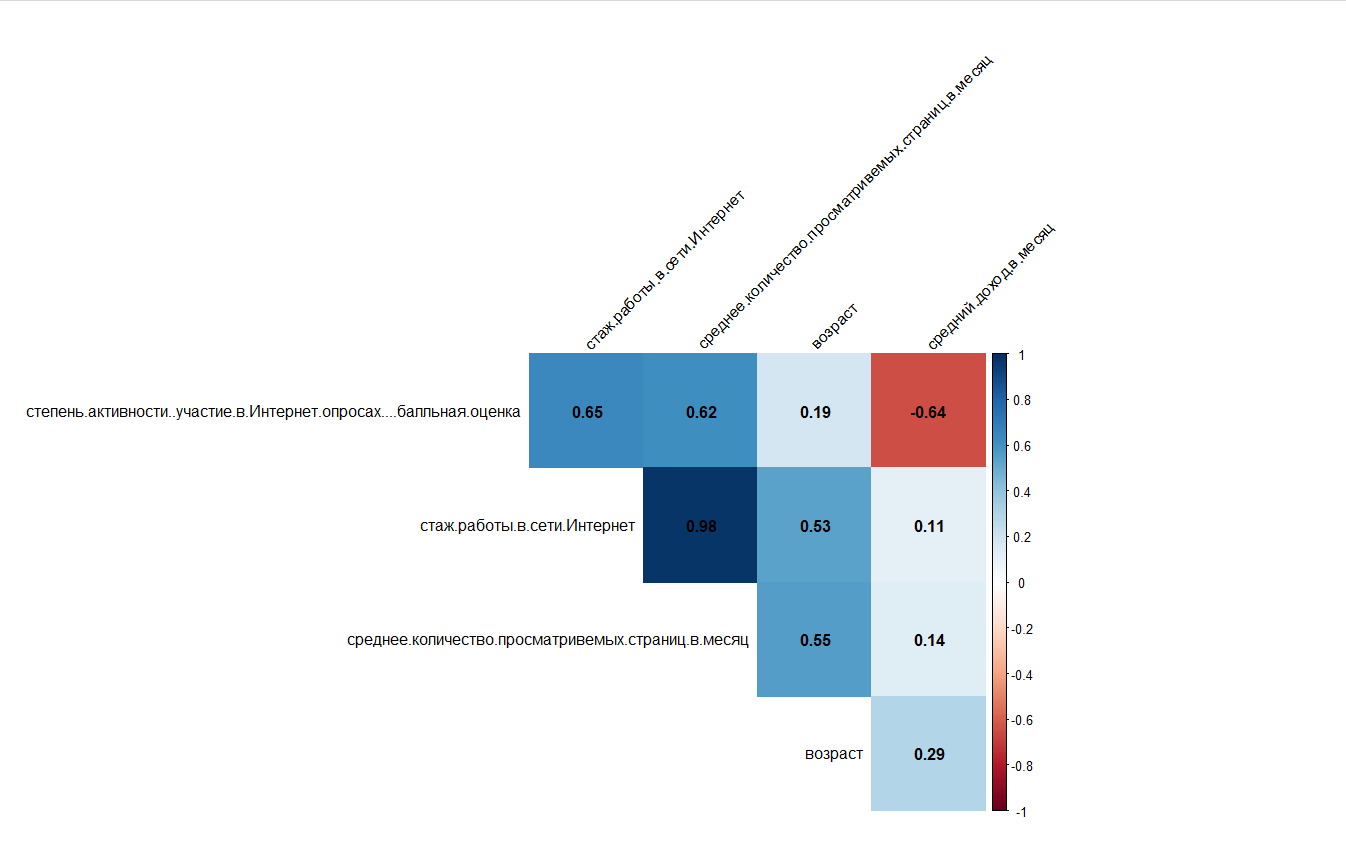


Рисунок 28. Корреляция методом Пирсона для второй группы

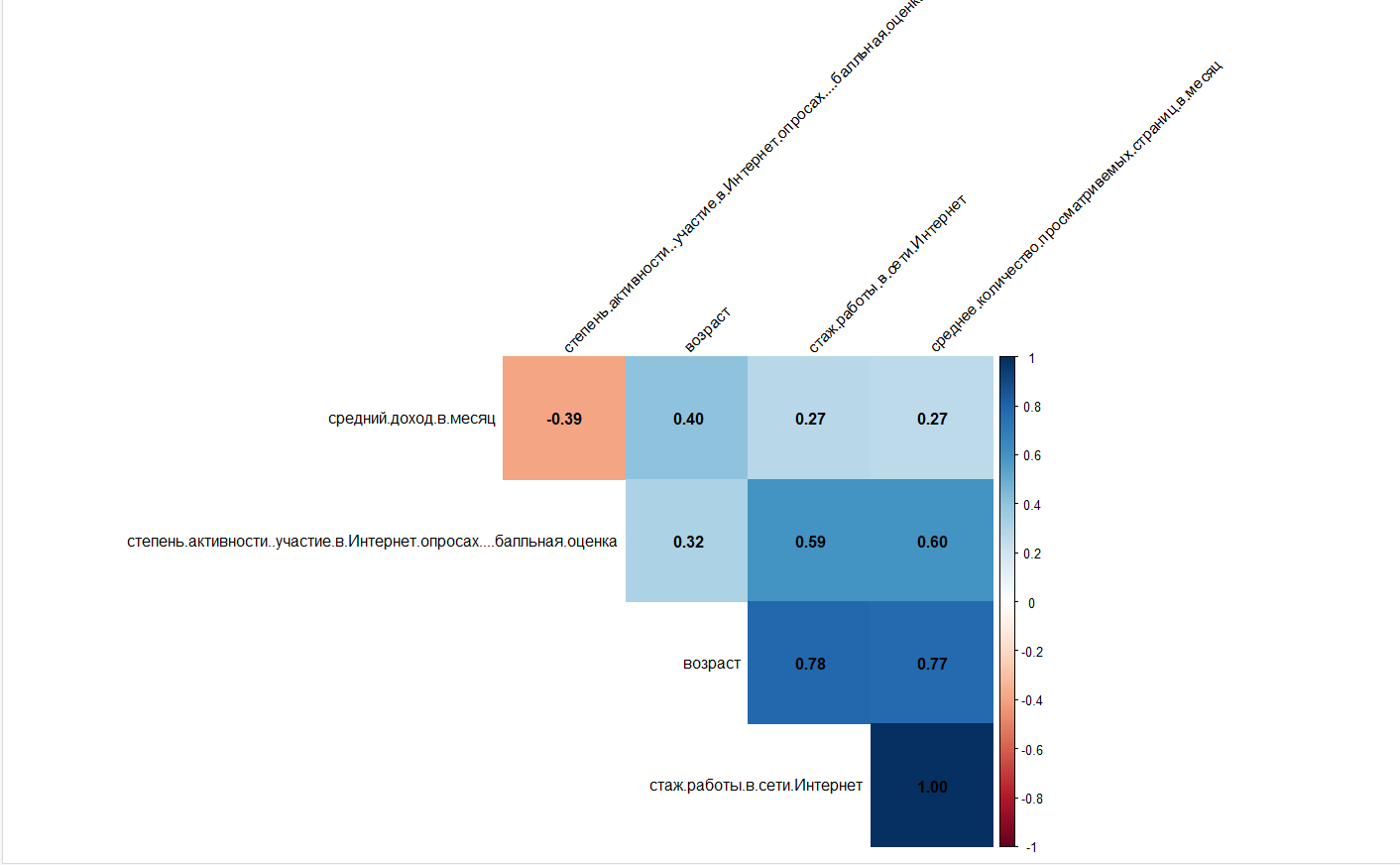


Рисунок 29. Корреляция методом Спирмена для первой группы

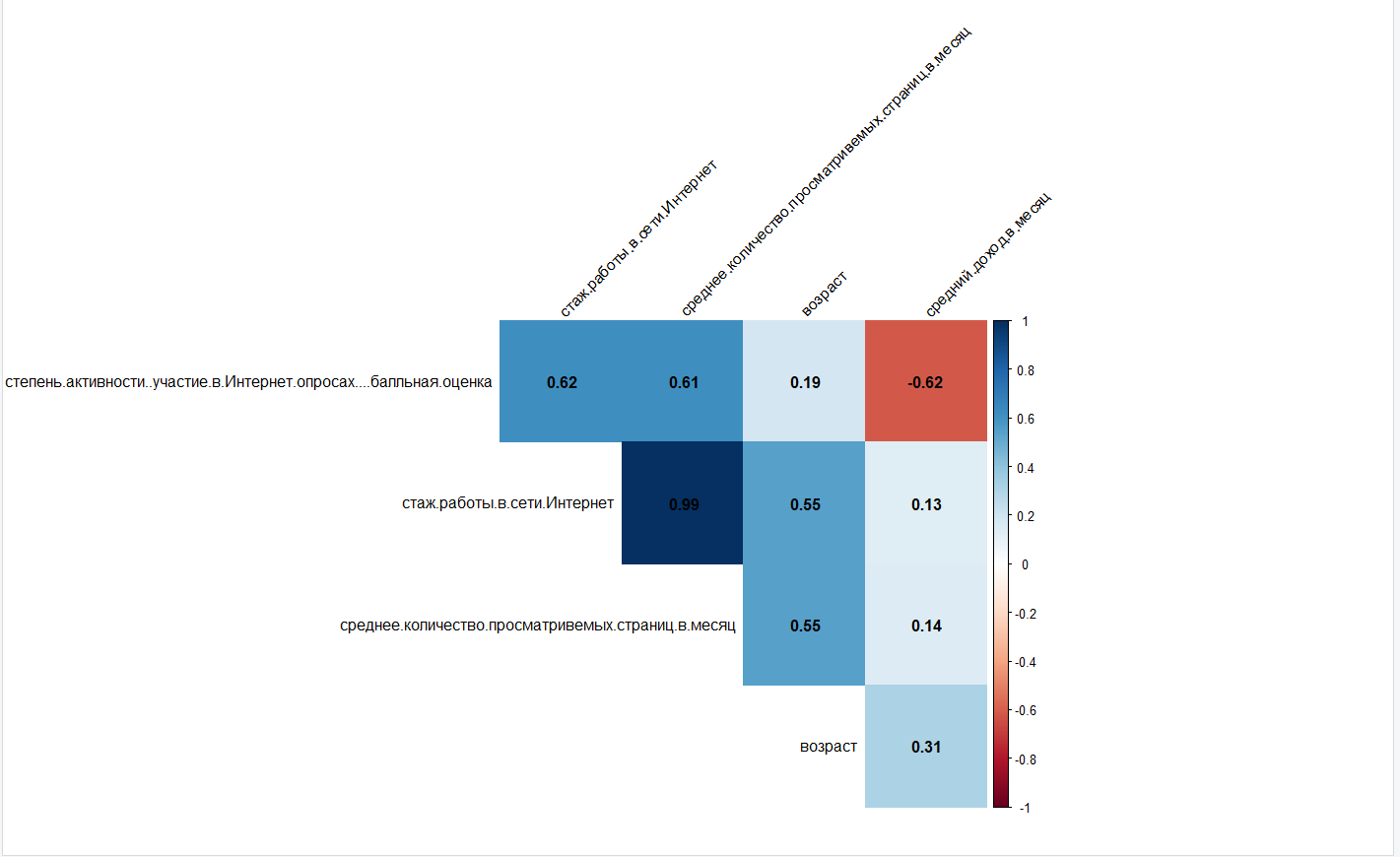


Рисунок 30. Корреляция методом Спирмена для второй группы

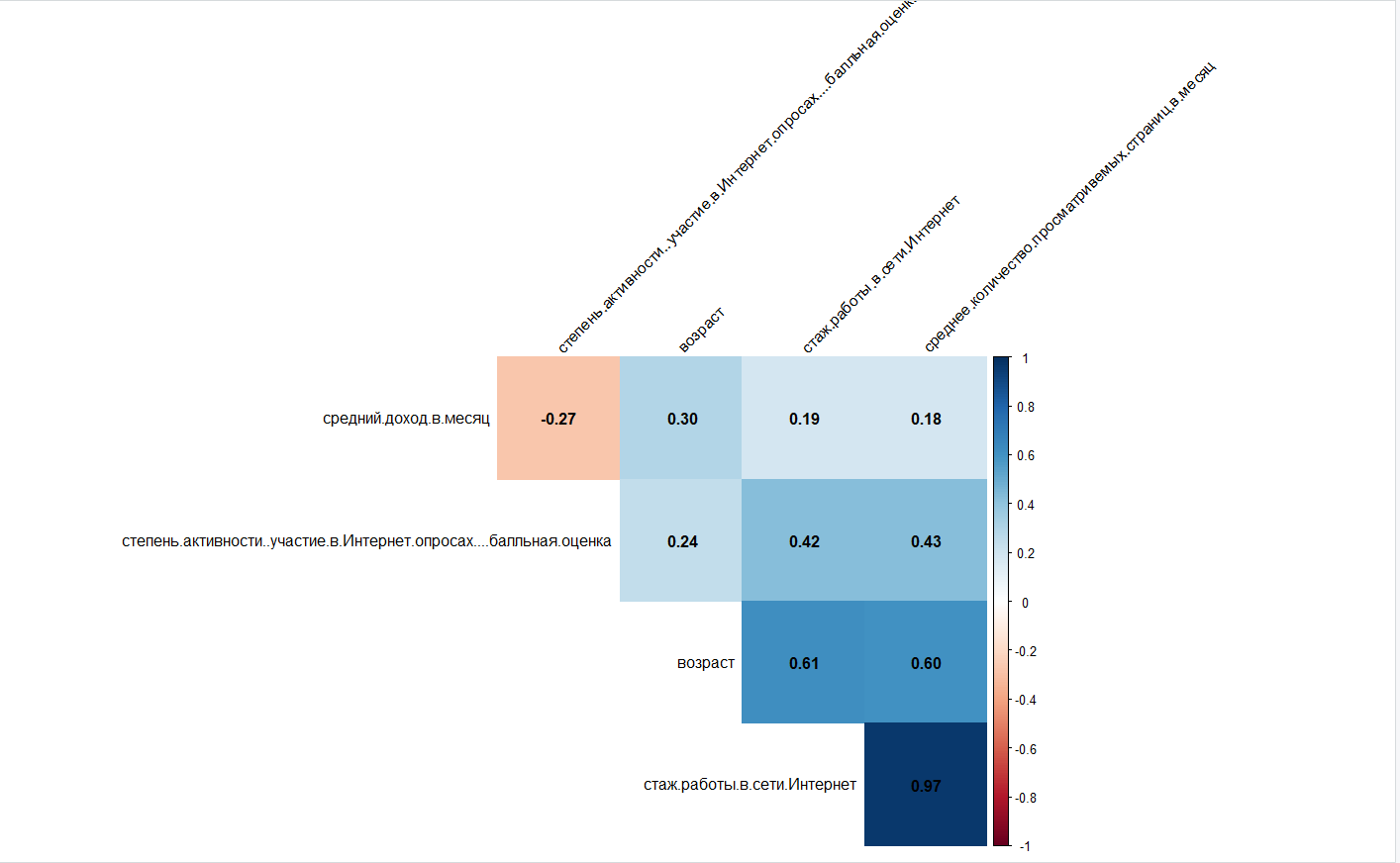


Рисунок 31. Корреляция методом Кендалла для первой группы

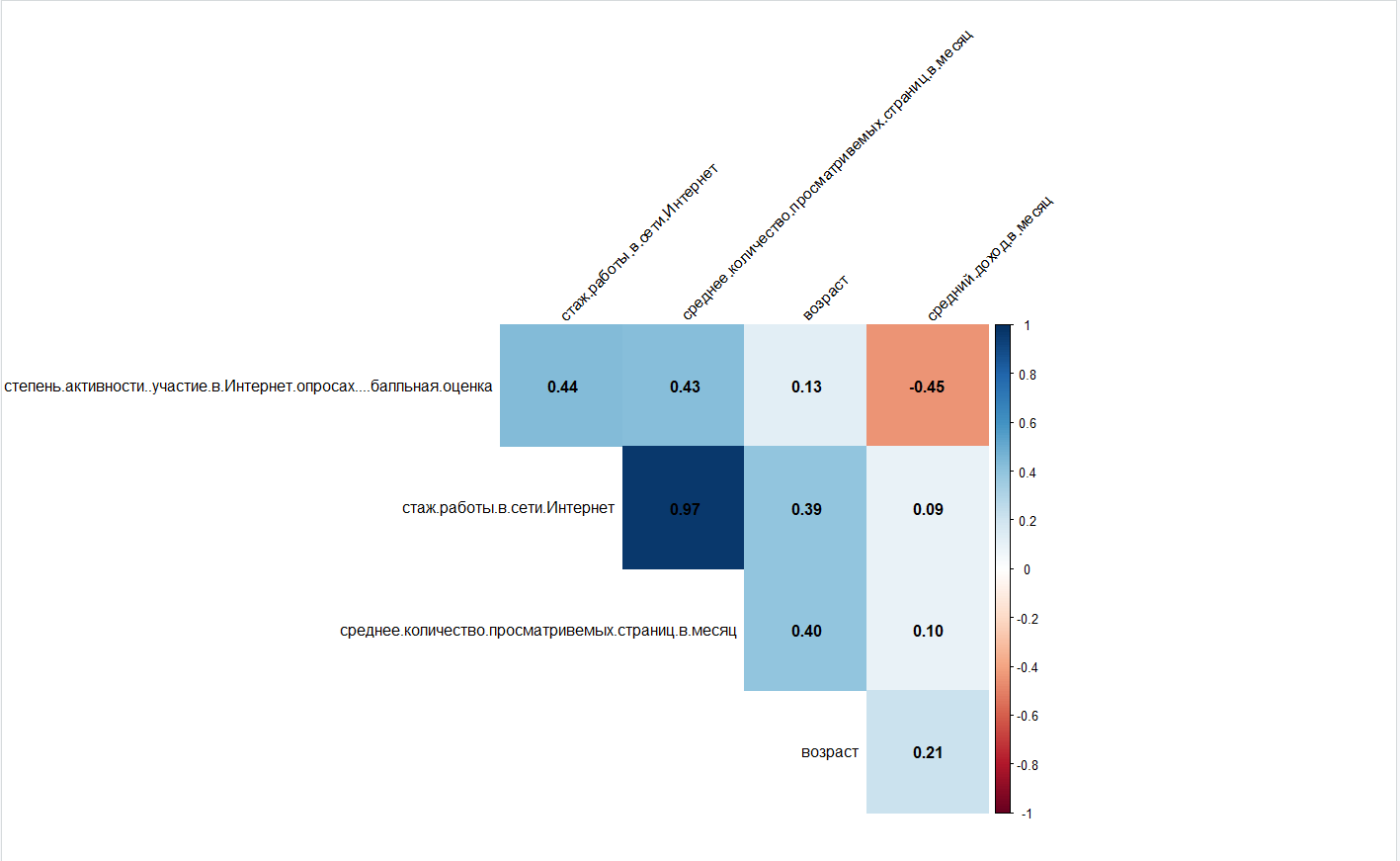


Рисунок 32. Корреляция методом Кендалла для второй группы

На основе этих данных можно сделать выводы:

* Средний доход в месяц: небольшая отрицательная корреляция со степенью активности в участии в интернет-опросах. Это указывает на то, что люди с более высоким доходом реже участвуют в опросах. Положительная корреляция с возрастом, указывает, что более возрастные участники получают больший доход.
* Степень активности в участии в интернет-опросах: Положительная корреляция с возрастом, а также с опытом работы в интернете. Указывает, что более возрастные люди с большим опытом в интернете более активно участвуют в опросах.
* Сильная положительная корреляция с опытом работы в интернете. Это указывает что старшие пользователи, вероятно, уже длительное время используют интернет.
* Среднее количество просматриваемых страниц в месяц: Положительная корреляция с возрастом и стажем работы в интернете, что может указывать на то, что старшие пользователи с большим опытом в интернете просматривают больше страниц.

4. Расчет частного коэффициента корреляции для первой и второй группы: будет рассматриваться стаж работы в сети интернет и среднее количество просматриваемых страниц в месяц, так как их коэффициент по модулю самый большой.

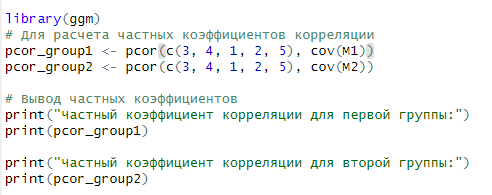


Рисунок 33. Частный коэффициент корреляции

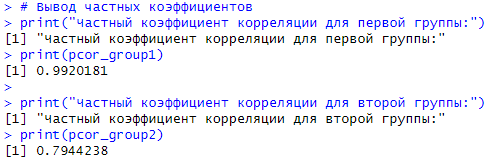


Рисунок 34. Результаты расчета

На рисунке 34 показаны результаты расчета частного коэффициента корреляции. У первой группы очень высокая положительная корреляция, это указывает на то, что между стаж работы в сети интернет и средним количеством просматриваемых страниц в месяц есть сильная связь. Во второй группе корреляция менее выражена.

5. Графическое представление матрицы коэффициентов корреляции.



Рисунок 35. Графическое представление матрицы

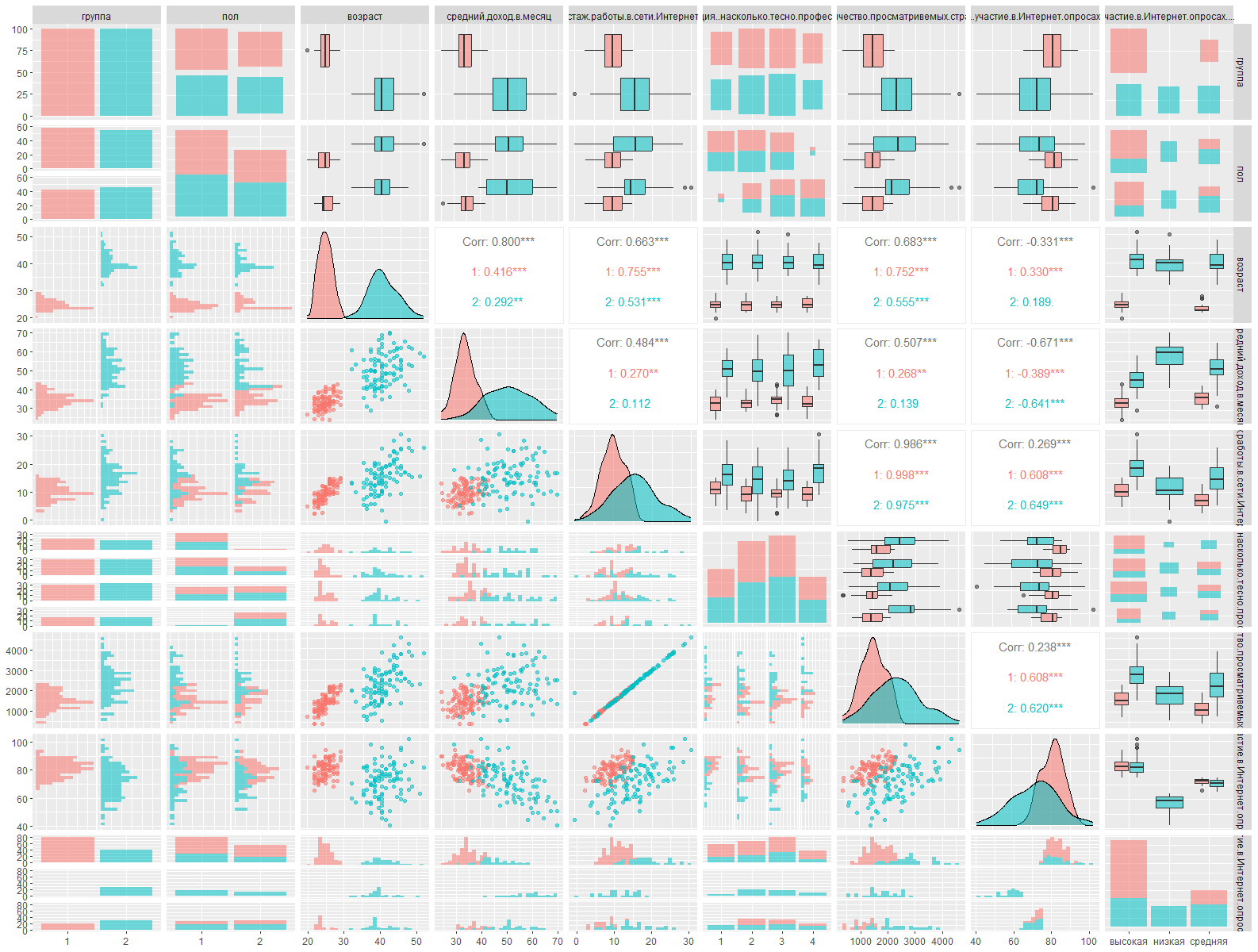


Рисунок 36. Графическое представление матрицы

**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:

* Средний доход в целом увеличивается с возрастом
* Стаж работы в сети интернет и среднее количество просматриваемых страниц в месяц имею сильную связь
* Присутствует значительная связь между полом и профессиональной специализацией
* Значительная часть людей имеет относительно небольшой стаж - от 0 до 10 лет. Стаж более 20–25 лет встречается гораздо реже.

**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)

# меняем с int на char

data$X.п.п <- as.character(data$X.п.п)

data$группа <- as.character(data$группа)

data$пол <- as.character(data$пол)

data$профессиональная.специализация..насколько.тесно.профессия.клиента.связана.с.Интернет. <- as.character(data$профессиональная.специализация..насколько.тесно.профессия.клиента.связана.с.Интернет.)

# 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)

pie(survey\_values,

labels = survey\_labels,

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

col = rainbow(length(survey\_values)))

# Категориальная радиальная диаграмма

# Установка графического окна для 2 строк и 3 столбцов (всего 6 диаграмм)

par(mfrow = c(2, 3))

# 1. Степень активности по полу

activity\_by\_gender <- table(data$степень.активности..участие.в.Интернет.опросах....качественная.оценка, data$пол)

# Для мужчин (пол = 1)

activity\_male <- activity\_by\_gender[, "1"]

labels\_male <- names(activity\_male)

pie(activity\_male, labels = labels\_male, main = "Степень активности - Мужчины", col = rainbow(length(activity\_male)), border = "white")

# Для женщин (пол = 2)

activity\_female <- activity\_by\_gender[, "2"]

labels\_female <- names(activity\_female)

pie(activity\_female, labels = labels\_female, main = "Степень активности - Женщины", col = rainbow(length(activity\_female)), border = "white")

# 2. Группа по полу

group\_by\_gender <- table(data$группа, data$пол)

# Для группы 1 (мужчины и женщины)

group\_male <- group\_by\_gender[,"1"]

group\_female <- group\_by\_gender[,"2"]

# Круговая диаграмма для мужчин по группе

pie(group\_male, labels = names(group\_male), main = "Группа - Мужчины", col = c("lightblue", "lightgreen"), border = "white")

# Круговая диаграмма для женщин по группе

pie(group\_female, labels = names(group\_female), main = "Группа - Женщины", col = c("lightblue", "lightgreen"), border = "white")

# 3. Профессиональная специализация по полу

specialization\_by\_gender <- table(data$профессиональная.специализация..насколько.тесно.профессия.клиента.связана.с.Интернет., data$пол)

# Для мужчин (пол = 1)

specialization\_male <- specialization\_by\_gender[, "1"]

labels\_male\_specialization <- names(specialization\_male)

pie(specialization\_male, labels = labels\_male\_specialization, main = "Специализация - Мужчины", col = rainbow(length(specialization\_male)), border = "white")

# Для женщин (пол = 2)

specialization\_female <- specialization\_by\_gender[, "2"]

labels\_female\_specialization <- names(specialization\_female)

pie(specialization\_female, labels = labels\_female\_specialization, main = "Специализация - Женщины", col = rainbow(length(specialization\_female)), border = "white")

par(mfrow = c(1, 1))

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

# Настройка для отображения 2 строк и 2 столбцов графиков

par(mfrow = c(2, 2))

# 1. Средний доход по полу (мужчины, женщины)

average\_income\_by\_gender <- tapply(data$средний.доход.в.месяц, data$пол, mean, na.rm = TRUE)

barplot(average\_income\_by\_gender,

beside = TRUE,

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

names.arg = c("Мужчины", "Женщины"),

main = "Средний доход по полу",

ylab = "Средний доход в месяц",

ylim = c(0, max(average\_income\_by\_gender)), # Для более четкого отображения

border = "white")

# 2. Средний доход по группе (1, 2)

average\_income\_by\_group <- tapply(data$средний.доход.в.месяц, data$группа, mean, na.rm = TRUE)

barplot(average\_income\_by\_group,

beside = TRUE,

col = c("lightgreen", "lightcoral"),

names.arg = c("Группа 1", "Группа 2"),

main = "Средний доход по группе",

ylab = "Средний доход в месяц",

ylim = c(0, max(average\_income\_by\_group)),

border = "white")

# 3. Средний доход по профессиональной специализации (1, 2, 3, 4)

average\_income\_by\_specialization <- tapply(data$средний.доход.в.месяц,

data$профессиональная.специализация..насколько.тесно.профессия.клиента.связана.с.Интернет.,

mean, na.rm = TRUE)

barplot(average\_income\_by\_specialization,

beside = TRUE,

col = c("lightblue", "lightgreen", "lightyellow", "lightpink"),

names.arg = c("Специализация 1", "Специализация 2", "Специализация 3", "Специализация 4"),

main = "Средний доход по специализации",

ylab = "Средний доход в месяц",

ylim = c(0, max(average\_income\_by\_specialization)),

border = "white")

# 4. Средний доход по степени активности в опросах (высокая, низкая, средняя)

average\_income\_by\_activity <- tapply(data$средний.доход.в.месяц,

data$степень.активности..участие.в.Интернет.опросах....качественная.оценка,

mean, na.rm = TRUE)

barplot(average\_income\_by\_activity,

beside = TRUE,

col = c("lightblue", "lightgreen", "lightcoral"),

names.arg = c("высокая", "средняя", "низкая"),

main = "Средний доход по активности",

ylab = "Средний доход в месяц",

ylim = c(0, max(average\_income\_by\_activity)),

border = "white")

# Возвращаем настройку графиков на стандартное отображение (1 график)

par(mfrow = c(1, 1))

# Диаграмма размаха для возраста в зависимости от пола

par(mfrow = c(2, 2))

boxplot(возраст ~ пол,

data = data,

main = "Диаграмма размаха: Возраст от пола",

ylab = "Возраст",

xlab = "Пол",

col = c("lightblue", "lightpink"), # Цвета для мужчин и женщин

border = "black")

# Диаграмма размаха для возраста в зависимости от группы (1, 2)

boxplot(возраст ~ группа,

data = data,

main = "Диаграмма размаха: Возраст от группы",

ylab = "Возраст",

xlab = "Группа",

col = c("lightgreen", "lightyellow"), # Цвета для группы 1 и 2

border = "black")

# Диаграмма размаха для возраста в зависимости от профессиональной специализации

boxplot(возраст ~ профессиональная.специализация..насколько.тесно.профессия.клиента.связана.с.Интернет.,

data = data,

main = "Диаграмма размаха: Возраст от специализации",

ylab = "Возраст",

xlab = "Профессиональная специализация",

col = rainbow(4), # 4 цвета для 4 специализаций

border = "black")

# Диаграмма размаха для возраста в зависимости от степени активности в опросах

boxplot(возраст ~ степень.активности..участие.в.Интернет.опросах....качественная.оценка,

data = data,

main = "Диаграмма размаха: Возраст от степени активности",

ylab = "Возраст",

xlab = "Степень активности",

col = c("lightblue", "lightgreen", "lightcoral"), # Цвета для высокой, средней и низкой активности

border = "black")

par(mfrow = c(1, 1))

# Настройка области графиков для отображения пяти графиков (2 строки и 3 столбца)

par(mfrow = c(2, 3)) # 2 строки и 3 столбца

# Гистограммы для количественных признаков

# Гистограмма для возраста

hist(data$возраст,

main = "Гистограмма возраста",

xlab = "Возраст",

col = "lightblue",

breaks = 10)

# Гистограмма для среднего дохода в месяц

hist(data$средний.доход.в.месяц,

main = "Гистограмма дохода",

xlab = "Средний доход в месяц",

col = "lightgreen",

breaks = 10)

# Гистограмма для стажа работы

hist(data$стаж.работы.в.сети.Интернет,

main = "Гистограмма стажа работы",

xlab = "Стаж работы",

col = "lightcoral",

breaks = 10)

# Гистограмма для количества просмотренных страниц

hist(data$среднее.количество.просматривемых.страниц.в.месяц,

main = "Гистограмма просмотров",

xlab = "Количество просмотренных страниц",

col = "lightyellow",

breaks = 10)

# Гистограмма для степени активности (бальная)

hist(data$степень.активности..участие.в.Интернет.опросах....балльная.оценка,

main = "Гистограмма активности (бальная)",

xlab = "Степень активности",

col = "lightpink",

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 + среднее.количество.просматривемых.страниц.в.месяц + степень.активности..участие.в.Интернет.опросах....балльная.оценка,

data = data,

main = "Матричный график количественных переменных",

pch = 21,

bg = "lightblue",

col = "darkblue")

#Корреляционный анализ

# Фишера

# Переменные для анализа

var <- c('пол', 'степень.активности..участие.в.Интернет.опросах....качественная.оценка', 'профессиональная.специализация..насколько.тесно.профессия.клиента.связана.с.Интернет.')

# Функция для обработки данных и выполнения тестов

perform\_tests <- function(data\_group, var) {

for (i in 1:(length(var) - 1)) {

for (j in (i + 1):length(var)) {

# Выбираем пары переменных

var1 <- var[i]

var2 <- var[j]

# Создание таблицы сопряжённости

table\_data <- table(data\_group[[var1]], data\_group[[var2]])

# Выполнение тестов χ² и Фишера

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

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

# Вывод результатов

print(paste("Анализ переменных:", var1, "и", var2))

print("Тест χ²:")

print(chi\_test)

print("Тест Фишера:")

print(fisher\_test)

}

}

}

# Перебор по группам и выполнение тестов для первой и второй группы

print("Результаты для первой группы")

perform\_tests(data\_group1, var)

print("Результаты для второй группы")

perform\_tests(data\_group2, var)

# ANOVA и критерий Краскела-Уоллиса

anova\_result <- aov(возраст ~ пол, data = data)

summary(anova\_result)

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

summary(anova\_result)

anova\_result <- aov(стаж.работы.в.сети.Интернет ~ пол, data = data)

summary(anova\_result)

anova\_result <- aov(среднее.количество.просматривемых.страниц.в.месяц ~ пол, data = data)

summary(anova\_result)

anova\_result <- aov(степень.активности..участие.в.Интернет.опросах....балльная.оценка ~ пол, data = data)

summary(anova\_result)

kruskal\_test <- kruskal.test(возраст ~ пол, data = data)

print(kruskal\_test)

kruskal\_test <- kruskal.test(средний.доход.в.месяц ~ пол, data = data)

print(kruskal\_test)

kruskal\_test <- kruskal.test(стаж.работы.в.сети.Интернет ~ пол, data = data)

print(kruskal\_test)

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

print(kruskal\_test)

kruskal\_test <- kruskal.test(степень.активности..участие.в.Интернет.опросах....балльная.оценка ~ пол, data = data)

print(kruskal\_test)

# Загрузка необходимых библиотек

library(ggm)

library(corrplot)

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

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

# Создание таблиц с числовыми переменными для первой и второй групп

M1 <- data\_group1[, unlist(lapply(data\_group1, is.numeric))]

M2 <- data\_group2[, unlist(lapply(data\_group2, is.numeric))]

# Коэффициенты корреляции для первой группы

N1\_group1 <- cor(M1, use="pairwise.complete.obs") # Коэффициенты Пирсона

N2\_group1 <- cor(M1, use="pairwise.complete.obs", method="spearman") # Коэффициенты Спирмена

N3\_group1 <- cor(M1, use="pairwise.complete.obs", method="kendall") # Коэффициенты Кендалла

# Коэффициенты корреляции для второй группы

N1\_group2 <- cor(M2, use="pairwise.complete.obs") # Коэффициенты Пирсона

N2\_group2 <- cor(M2, use="pairwise.complete.obs", method="spearman") # Коэффициенты Спирмена

N3\_group2 <- cor(M2, use="pairwise.complete.obs", method="kendall") # Коэффициенты Кендалла

# Вывод коэффициентов корреляции

print("Корреляция Пирсона для первой группы")

print(N1\_group1)

print("Корреляция Спирмена для первой группы")

print(N2\_group1)

print("Корреляция Кендалла для первой группы")

print(N3\_group1)

print("Корреляция Пирсона для второй группы")

print(N1\_group2)

print("Корреляция Спирмена для второй группы")

print(N2\_group2)

print("Корреляция Кендалла для второй группы")

print(N3\_group2)

# Графическое представление для первой группы

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))

corrplot(N1\_group1, method="color", col=NULL,

type="upper", order="hclust",

addCoef.col = "black", tl.col="black", tl.srt=45,

sig.level = 0.01, insig = "blank", diag=FALSE)

# Графическое представление для второй группы

corrplot(N1\_group2, method="color", col=NULL,

type="upper", order="hclust",

addCoef.col = "black", tl.col="black", tl.srt=45,

sig.level = 0.01, insig = "blank", diag=FALSE)

library(ggm)

# Для расчета частных коэффициентов корреляции

pcor\_group1 <- pcor(c(3, 4, 1, 2, 5), cov(M1))

pcor\_group2 <- pcor(c(3, 4, 1, 2, 5), cov(M2))

# Вывод частных коэффициентов

print("Частный коэффициент корреляции для первой группы:")

print(pcor\_group1)

print("Частный коэффициент корреляции для второй группы:")

print(pcor\_group2)

library(GGally)

ggpairs(data, columns = 2:10, aes(color = группа,alpha = 0.5))