ОТЧЕТ  
ПО ЛАБОРАТОРНОЙ РАБОТЕ №1  
ПО ДИСЦИПЛИНЕ «МЕТОДЫ АНАЛИЗА ДАННЫХ»

Факультет: АВТ Преподаватель:

Группа: АВТ-219 Зеленчук Н.А.

Студенты:

Димаков А.В.

Заверткин М.А.

Лысак А.Д.

Вариант 1

# **Введение в R. Методы первичного разведочного анализа данных в R**

**1. Цель лабораторной работы**

* 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).

**2. Задание к лабораторной работе**

Прочитайте содержательную постановку задачи для вашего варианта. Выберите файл с данными (в формате .csv), соответствующий вашему варианту.

1. Загрузите файл с данными, соответствующий варианту.
2. Просмотрите загруженную таблицу с данными. Попробуйте использовать команды R для доступа к определенному столбцу/строке данных, редактирования данных, получения подвыборок из данных в соответствии с заданными условиями.
3. Посмотрите структуру данных. Рассчитайте основные статистические характеристики по количественным данным (минимальное, максимальное, среднее значение, стандартное отклонение, первый и третий квартили, медиана, мода, асимметрия, эксцесс) отдельно для первой и второй групп и для всей выборки. Сделайте выводы.
4. Проведите графический анализ данных, постройте:

- диаграмму рассеяния по двум количественным признакам;

- радиальную диаграмму по качественному признаку;

- категориальную радиальную диаграмму по одному из качественных признаков в зависимости от пола и группы;

- категориальную столбиковую диаграмму по одному из количественных признаков в зависимости от пола или группы;

- диаграмму размаха для одного из количественных признаков в зависимости от значений пола или группы;

- гистограммы для всех количественных признаков на одном графике;

- матричный график по всем количественным переменным.

На основе проведенного анализа сделайте выводы о структуре данных, о характере распределения данных в терминах решаемой задачи.

1. Проверьте гипотезу о соответствии данных модели нормального закона распределения для одного из количественных признаков (отдельно для первой и второй группы) на основе критериев Шапиро – Уилка, Крамера – Мизеса, Андерсона – Дарлинга. Cделайте выводы в терминах прикладной задачи.
2. Проведите корреляционный анализ данных.

6.1. Оцените степень взаимосвязи между качественными переменными на основе критериев χ2 (Chi-квадрат) и Фишера для первой и второй групп. Сделайте выводы о силе и направлении связи в терминах решаемой задачи.

6.2. Оцените степень взаимосвязи между одной из качественных переменных и количественными переменными на основе использования однофакторного дисперсионного анализа (ANOVA) и критерия Краскела-Уоллиса.

6.3. Оцените степень взаимосвязи между количественными переменными на основе расчета коэффициентов корреляции Пирсона, Спирмена, Кендалла для первой и второй групп.

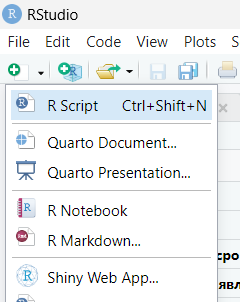
6.4. Оцените степень взаимосвязи между двумя количественными переменными (для которых коэффициент корреляции Пирсона максимален по модулю) на основе расчета частного коэффициента корреляции для первой и второй групп.

6.5. Графически представьте матрицы коэффициентов корреляции. Оцените статистическую значимость связи. Сделайте выводы о силе и направлении связи в терминах решаемой задачи. Постройте матричный график с помощью пакета ggpairs.

7. Сделайте выводы по работе в произвольной форме.

**3. Ход работы**

1. Установим рабочую директорию. Создадим файл скрипта.

****

*Рис.1 Установка рабочей директории*

Загрузим файл с данными, соответствующий варианту 2 с помощью команды:

data <- read.table("C:\\Users\\Downloads\\data.csv", header=TRUE, sep=";")

Для корректного отображения подключим русский язык:

Sys.setlocale("LC\_CTYPE", "russian")

А также переименуем переменные для удобного отображения страницы:

colnames(data)[5] <- 'стаж'

colnames(data)[6] <- 'процент.выпол.разраб.в\_срок'

colnames(data)[7] <- 'колич.ошиб.от\_пользов.'

colnames(data)[8] <- 'бальн.оцен.заказчика'

colnames(data)[9] <- 'качеств.оцен.заказчика'

colnames(data)[10] <- 'кач.документир.'

1. Рассчитаем основные статистические характеристики по количественным данным, используя пакет psych и его функцию describe().

Так как некоторые характеристики не являются количественными, преобразуем их тип в factor, чтобы они не использовались в расчете статистики.

data$группа <- as.factor(data$группа)

data$пол <- as.factor(data$пол)

data$качеств.оцен.заказчика <- as.factor(data$качеств.оцен.заказчика)

Разделим общую таблицу на две таблицы — для групп 1 и 2 для количественных и качественных:

tableNum1 <- subset(data,группа==1,select=c("группа","возраст","стаж","процент.выпол.разраб.в\_срок","колич.ошиб.от\_пользов.","бальн.оцен.заказчика","кач.документир."))

View(tableNum1)

tableNum2 <- subset(data,группа==2,select=c("группа","возраст","стаж","процент.выпол.разраб.в\_срок","колич.ошиб.от\_пользов.","бальн.оцен.заказчика","кач.документир."))

View(tableNum2)

tableCat1 <- subset(data,группа==1, select=c("группа", "пол", "качеств.оцен.заказчика"))

View(tableCat1)

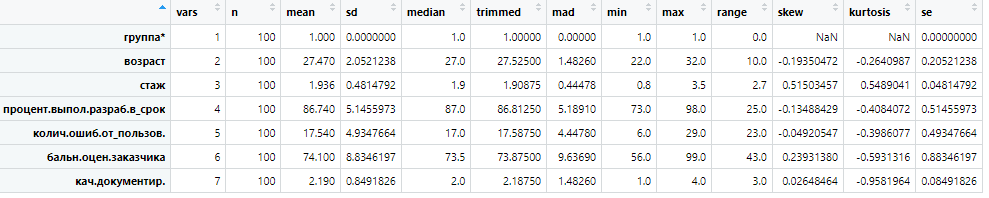
tableCat2 <- subset(data,группа==2, select=c("группа", "пол", "качеств.оцен.заказчика"))

View(tableCat2)

Рассчитаем характеристики для первой группы:РРрр

tableNumDescribed1 <- describe(tableNum1)

View(tableNumDescribed1)

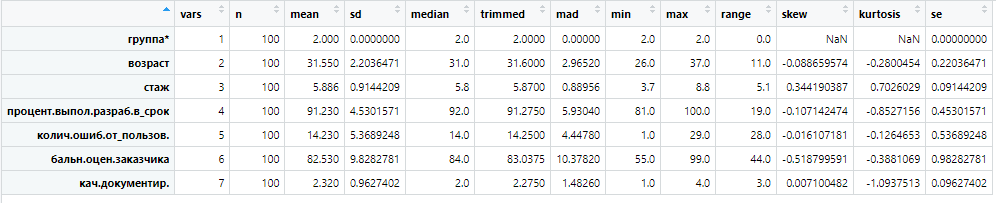


*Рис. 2 Характеристики для первой группы*

Для второй группы:

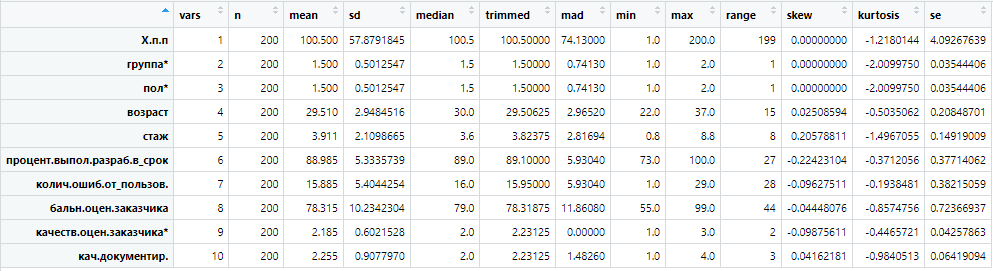
tableNumDescribed2 <- describe(tableNum2)

View(tableNumDescribed2)



*Рис. 3 Характеристики для второй группы*

Для обеих групп:



*Рис. 4 Характеристики для обеих групп*

Здесь:

Mean — медиана

Sd — среднеквадратичное отклонение

Min — минимальное значение

Max — максимальное значение

Range — разница между мин и макс (размах)

Skew — асимметрия

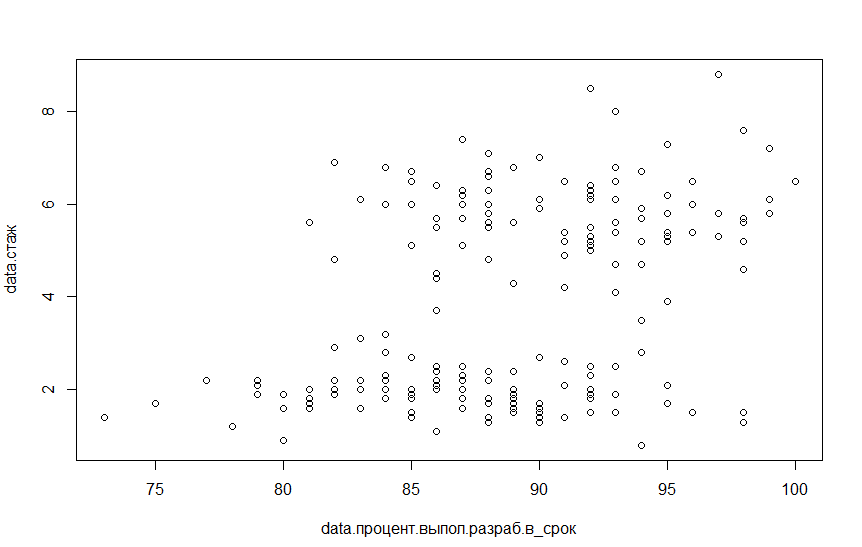
Kurtosis — эксцесс

Также, некоторые характеристики можно рассчитать с помощью функции summary()

1. Проведем графический анализ данных.

* Построим диаграмму рассеяния по двум количественным признакам — выполненной разработки в срок (процент) и стажу:

plot(data.frame(data$процент.выпол.разраб.в\_срок, data$стаж))



*Рис.5 Диаграмма рассеяния по выполненной разработки в срок (процент) и стажу*

**Вывод:**

1. **Общая тенденция**: Можно отметить слабую положительную тенденцию — в среднем у сотрудников с большим стажем наблюдаются немного более высокие значения процента выполнения, но это не однозначно и выражено слабо.
2. **Распределение точек**: данные образуют по крайней мере два видимых кластера по стажу: один — низкий стаж (примерно ~1–3 года), другой — средний/высокий (примерно ~5–7+ лет). Оба кластера покрывают широкий диапазон значений процента выполнения (то есть и новички, и опытные работники встречаются как с низкими, так и с высокими процентами выполнения).
3. **Широкий диапазон в старших возрастах**: при высоких процентах выполнения: при значениях **процент выполнения** примерно от ~90 до 100 разброс стажа заметно больший — там присутствуют и малый, и большой стаж. Это указывает, что высокий процент выполнения достижим для сотрудников с разным опытом; стаж сам по себе не является единственным фактором, определяющим высокий процент выполнения.

Таким образом, данные демонстрируют, что стаж работы не является единственным определяющим фактором процента выполнения работ в срок: прямая зависимость, если и существует, — слабая, а разброс значений показывает, что высокий процент выполнения наблюдается у сотрудников с разным опытом.

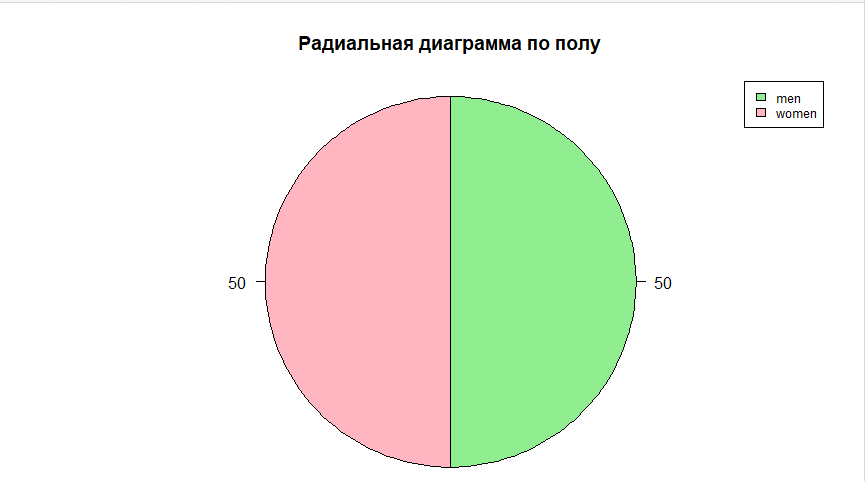
data1 <- data$пол

x <- c(summary(data1))

piepercent <- round(100\*x/sum(x),1)

pie(x,piepercent,radius=1, main="Радиальная диаграмма по полу", col=c("lightgreen","lightpink"), clockwise = TRUE)

legend("topright", c("men", "women"), cex=0.8,fill=c("lightgreen", "lightpink"))



*Рис.6 Радиальная диаграмма по полу*

Женщин и мужчин – одинаковое количество.

* Построим категориальную радиальную диаграмму по степени удовлетворенности заказчика в зависимости от пола и группы:

plot.new()

data1<-data$проф.связь

x1 <- c(summary(data1[data$пол==1&data$группа==1]))

x2 <- c(summary(data1[data$пол==1&data$группа==2]))

x3 <- c(summary(data1[data$пол==2&data$группа==1]))

x4 <- c(summary(data1[data$пол==2&data$группа==2]))

piepercent <- round(100\*x1/sum(x1),1)

piepercent <- round(100\*x2/sum(x2),1)

piepercent <- round(100\*x3/sum(x3),1)

piepercent <- round(100\*x4/sum(x4),1)

par (mfrow=c(2,2))

pie(x1,piepercent,radius=1, main="Радиальная диаграмма по полу и группе", xlab= "Муж", ylab="1", col=c("#149","#773049", "pink"), clockwise = TRUE)

legend("topright", c("men", "group"), cex=0.8,fill=c("#149", "#773049"))

pie(x2,piepercent,radius=1, main="Радиальная диаграмма по полу и группе", xlab= "Муж", ylab="2", col=c("#149","#494","pink"), clockwise = TRUE)

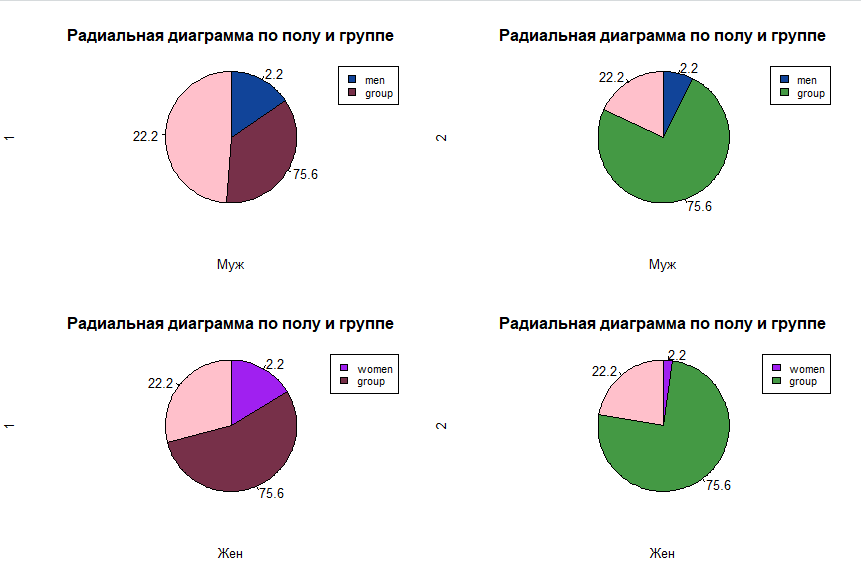
legend("topright", c("men", "group"), cex=0.8,fill=c("#149", "#494"))

pie(x3,piepercent,radius=1, main="Радиальная диаграмма по полу и группе", xlab= "Жен", ylab="1", col=c("purple","#773049","pink"), clockwise = TRUE)

legend("topright", c("women", "group"), cex=0.8,fill=c("purple","#773049"))

pie(x4,piepercent,radius=1, main="Радиальная диаграмма по полу и группе", xlab= "Жен", ylab="2", col=c("purple","#494","pink"), clockwise = TRUE)

legend("topright", c("women", "group"), cex=0.8,fill=c("purple","#494"))



*Рис.7 Категориальная радиальная диаграмма по степени удовлетворенности заказчика в зависимости от пола и группы*

* Мужчины и женщины в разных группах демонстрируют различия в распределении качественной оценки заказчика: доли категорий между полами не одинаковы и заметно отличаются визуально.
* В группе 1 наблюдается явное преимущество одной категории у женщин — одна секция диаграммы значительно больше остальных; аналогичная картина видна и для женщин в группе 2. Иначе: у женщин одна категория доминирует в обеих группах.
* У мужчин распределение по категориям выглядит более равномерным по сравнению с женщинами — у мужчин одна категория не так явно доминирует, доли распределены более сбалансировано.
* Визуально видно три уровня (очевидно: большая доля ≈75.6%, средняя ≈22.2%, малая ≈2.2%) — это указывает на сильную концентрацию ответов в одной категории и малые доли в остальных, особенно у женщин.
* Построим категориальную столбиковую диаграмму по стажу в зависимости от пола и группы:

data\_sorted <- with(data,data[order(возраст), ])

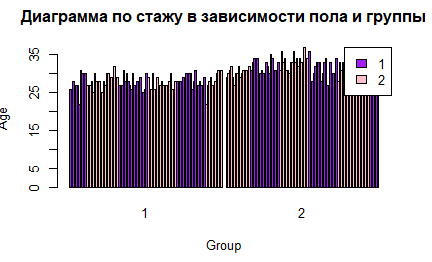
barplot

data\_grouped <- split(data\_sorted$возраст, list(data\_sorted$пол, data\_sorted$ группа))

matrixq <- matrix(unlist(data\_grouped), ncol=2)

barplot(matrixq,beside = TRUE, space =c (0,2.5), col=c("purple", "pink"), names.arg=unique(data\_sorted$группа), main ="Диаграмма по стажу в зависимости пола и группы", xlab="Group", ylab="Age")

legend("topright", legend=unique(data$группа), fill=c("purple","pink"))

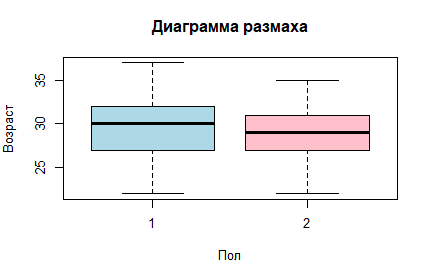


*Рис.8 Категориальная столбиковая диаграмма по стажу в зависимости от пола и группы*

Здесь можно выявить, что во второй группе средний стаж выше, чем в первой группе.

* Построим диаграмму размаха для возраста в зависимости от пола:

boxplot(возраст ~ пол, data = data, main = "Диаграмма размаха", xlab = "Пол", ylab = "Возраст", col = c("lightblue", "pink"))



*Рис.9 Диаграмма размаха для возраста в зависимости от пола*

**Вывод:**

1. **Общая тенденция.** Возраст у мужчин и женщин примерно одинаков — сильной разницы не видно. Медианы (средние «центры») обеих групп близки.
2. **Распределение.** Большая часть людей в обеих группах — примерно одного возраста (похоже на конец двадцатых — начало тридцатых лет). Интерквартильные интервалы (центральные 50%) у групп похожи по ширине.
3. **Разброс и выбросы.** У одной из групп верхняя граница немного выше — там есть несколько более старших людей. Явных больших выбросов нет, в целом разбросы сопоставимы.

Возраст мужчин и женщин в выборке существенно не различается — распределения похоже совпадают, разница небольшая.

* Теперь, построим гистограммы для всех количественных признаков на одном графике

hist(data$возраст, main = "Гистограммы всех количественных

признаков", col = "lightblue", xlab = "Значения", ylab = "Частота", xlim= c(0,

max(data$`актив.балл.оц`)))

hist(data$стаж.работы, col = "#469", add = TRUE)

hist(data$ср.доход, col =

"purple", add = TRUE)

hist(data$актив.балл.оц,

col = "lightyellow", add = TRUE)

# Добавим легенду

legend(x = "top", y = "center",

legend = c("возраст", "стаж.работы", "доход", "активность балл.оц"),

cex = 0.5, fill = c("lightblue", "#469", "purple","lightyellow"))

par(mfrow = c(1, 1))

# Создадим матричный график

pairs(data[, c("возраст", "стаж.работы",

"ср.кол.страниц" ,

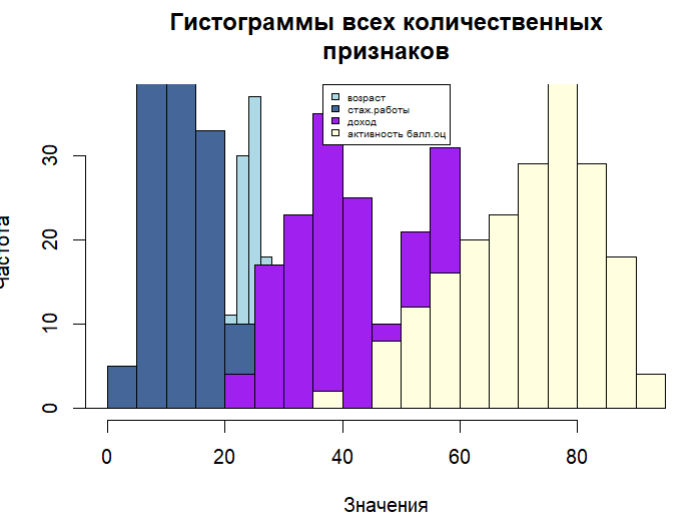
"актив.балл.оц")],

col = c("blue", "#127", "#699", "purple", "#774"),

main = "Матричный график всех количественных признаков",

pch = 21, bg = "grey", # Параметры точек (фон белый)

cex = 0.5) # Уменьшаем размер точек



*Рис.10 Гистограммы для всех количественных признаков*

* Создадим матричный график по всем количественным переменным:

par(mfrow = c(1, 1))

# Создадим матричный график

pairs(data[, c("возраст", "стаж.работы",

"Выполнения\_разработок\_в\_срок" ,

"степень.уд-ния.заказч.балл.оц")],

col = c("blue", "coral", "aquamarine", "purple", "lavender"),

main = "Матричный график всех количественных признаков",

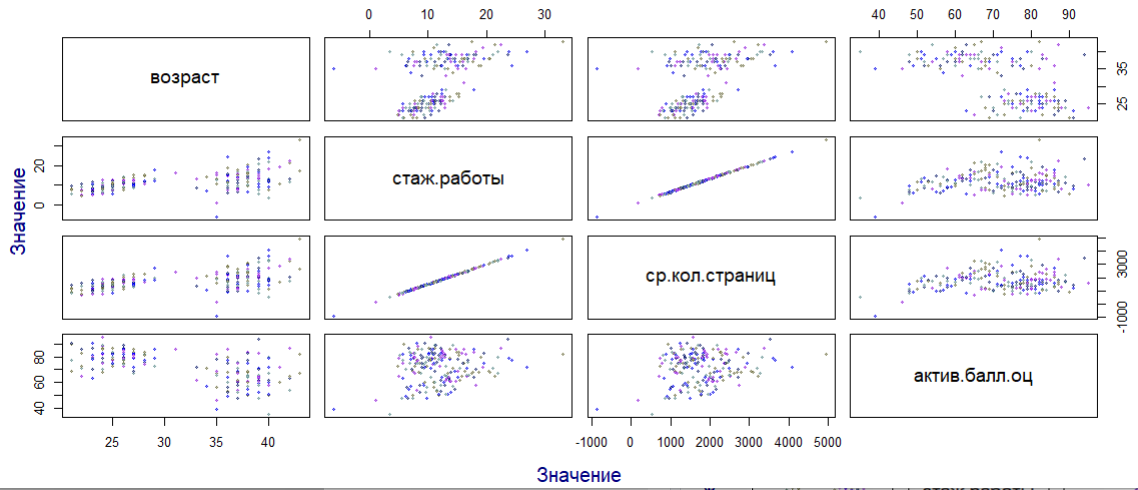
pch = 21, bg = "grey", # Параметры точек (фон белый)

cex = 0.5) # Уменьшаем размер точек

# Добавим подписи осей

mtext("Значение", side = 1, line = 4, col = "navy", cex = 1) #Подпись оси X

mtext("Значение", side = 2, line = 2.5, col = "navy", cex = 1) #Подпись оси Y



*Рис. 11 Матричный график по всем количественным переменным*

**Вывод:**

1. **Взаимосвязь между возрастом и стажем работы**: Четко прослеживается положительная линейная зависимость, что логично, так как с увеличением возраста увеличивается и стаж.
2. **Взаимосвязь между стажем работы и средним количеством страниц**: Видна сильная положительная корреляция, указывающая, что люди с большим стажем обычно обрабатывают больше страниц.
3. **Взаимосвязь между средним количеством страниц и активными баллами**: Присутствует тенденция к положительной корреляции, что может означать, что большее количество страниц связано с более высокой активностью.
4. **Взаимосвязь между возрастом и активными баллами**: Корреляция менее выражена, но можно заметить некоторую зависимость, что может свидетельствовать о влиянии возраста на активность.
5. **Разные распределения**: Видны отличия в плотности точек для различных переменных, что указывает на различное распределение значений по переменным.

В целом, основные зависимости хорошо выражены для пары "возраст – стаж работы" и "стаж работы – среднее количество страниц", что может свидетельствовать о значительном влиянии этих факторов друг на друга.

1. Проведем корреляционный анализ данных.
   1. Оценим степень взаимосвязи между качественными переменными на основе критериев χ2 (Chi-квадрат) и Фишера для первой и второй групп.

**Тест Хи-квадрат**

R-код для теста Хи-квадрат:

chisq1 <- chisq.test(table(tl1$пол,tl1$актив.кач.оц))

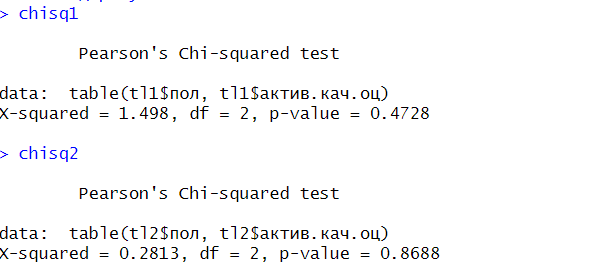
chisq2 <- chisq.test(table(tl2$пол,tl2$актив.кач.оц))

# Вывод результатов теста

chisq1

chisq2

Результат:



**Результаты теста хи-квадрат**

1. **chisq1:**
   * Значение хи-квадрат: 1.498
   * Степени свободы (df): 2
   * p-значение: 0.4728

**Интерпретация**: поскольку p-значение больше 0.05, мы не можем отвергнуть нулевую гипотезу об отсутствии связи между полом и качественной оценкой активности для данных tl1. Это указывает на то, что различия в оценках активности по полу могут быть случайными, и нет значимой статистической зависимости между этими переменными.

1. **chisq2:**
   * Значение хи-квадрат: 0.2813
   * Степени свободы (df): 2
   * p-значение: 0.8688

**Интерпретация**: здесь также p-значение намного больше 0.05. Таким образом, различия в оценках активности по полу в данных tl2 можно объяснить случайностью, и статистически значимой связи между полом и оценкой активности нет.

**Общий вывод:** В обоих случаях отсутствует статистически значимая связь между полом и оценкой активности. Это говорит о том, что пол не оказывает существенного влияния на оценки качественной активности в данных.

**Тест Фишера**

R-код для теста Фишера:

# Тест Фишера

fisher1 <- fisher.test(table(tl1$пол,

tl1$ актив.кач.оц))

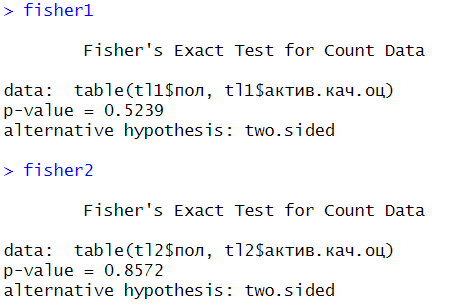
fisher2 <- fisher.test(table(tl2$пол,

tl2$ актив.кач.оц))

fisher1

fisher2

Результаты:



Тест Фишера используется для определения наличия статистически значимой связи между двумя категориальными переменными. Он особенно полезен, когда выборка небольшая, и точность оценок важна. Этот тест помогает избежать ложных выводов, которые могут возникнуть при использовании теста хи-квадрат.

* **fisher1:**
* p-значение: 0.5239

**Интерпретация:** так как p-значение больше 0.05, это подтверждает отсутствие статистически значимой связи между полом и качественной оценкой активности для данных tl1. Данные не дают оснований считать, что пол влияет на оценку активности.

* **fisher2:**
* p-значение: 0.8572

**Интерпретация:** тест Фишера для данных tl2 также показывает, что p-значение значительно выше 0.05, что свидетельствует об отсутствии значимой зависимости между полом и качественной оценкой активности. Это поддерживает вывод, что пол и оценка активности не связаны.

* 1. Оценим степень взаимосвязи между одной из качественных переменных и количественными переменными на основе использования однофакторного дисперсионного анализа (ANOVA).

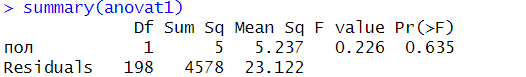
Для оценки степени взаимосвязи между качественной переменной и количественными переменными на основе однофакторного дисперсионного анализа (ANOVA), нужно выбрать одну качественную переменную и провести сравнение средних значений количественной переменной среди разных уровней качественной переменной.

После этого проводится однофакторный дисперсионный анализ (ANOVA). Результаты теста позволят оценить, является ли различие в средних значениях "стажа работы" статистически значимым с учетом разных "полов".

anovat1 <- aov(стаж.работы ~ пол, data = data)

summary(anovat1)

В коде мы используем функцию aov() для проведения однофакторного дисперсионного анализа. Зависимая переменная это стаж.работы, а независимая переменная - пол. Результаты анализа можно посмотреть с помощью команды summary(anova\_result).



Результаты следующие:

* + Степени свободы (Df) для пола: 1
  + Сумма квадратов (Sum Sq) для пола: 5.237
  + Средний квадрат (Mean Sq) для пола: 5.237
  + Значение статистики F: 0.226
  + p-значение (Pr(>F)): 0.635

1. Интерпретация:
   * Поскольку p-значение (0.635) значительно больше 0.05, мы не можем отвергнуть нулевую гипотезу, утверждающую, что стаж работы не зависит от пола. Это указывает на то, что разница в среднем стаже работы между различными полами статистически незначима.
   * Вклад переменной пол в объяснение вариации стажа работы очень мало (значение F низкое).

**Заключение**

Пол, вероятно, не оказывает значимого влияния на стаж работы в данной выборке. Выводы основаны на высоком p-значении, которое говорит о том, что различия в стаже между полами могут быть случайными.

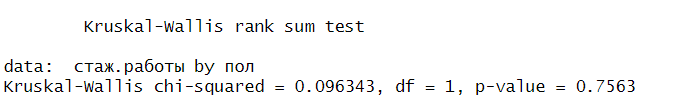
**Тест Краскала-Уоллиса**:

# Выполнение теста Краскела-Уоллиса

kruskal\_test\_result <- kruskal.test(стаж.работы ~ пол, data = tl1)

# Вывод результата теста

print(kruskal\_test\_result)



На основании результатов теста Краскала-Уоллиса:

1. **Тест Крускала-Уоллиса** используется для сравнения медиан нескольких групп, когда данные могут не соответствовать предположениям о нормальности (непараметрический тест). Он подходит, если данные являются порядковыми или распределены ненормально.
2. **Результаты:**
   * Хи-квадрат Крускала-Уоллиса: 0.096343
   * Степени свободы (df): 1
   * p-значение: 0.7563
3. **Интерпретация:**
   * p-значение (0.7563) значительно больше 0.05, что означает отсутствие статистически значимой разницы в распределении стажа работы между группами, разделенными по полу.
   * Нулевую гипотезу о том, что медианы стажа работы одинаковы для всех групп, нельзя отвергнуть.

**Заключение**

Согласно тесту Крускала-Уоллиса, различия в стаже работы между полами не являются значимыми. Это говорит о том, что распределение стажа работы в группах мужчин и женщин похоже.

* 1. Оценим степень взаимосвязи между количественными переменными на основе расчета коэффициентов корреляции Пирсона, Спирмена, Кендалла для первой и второй групп.

# Вывод результатов корреляций

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)

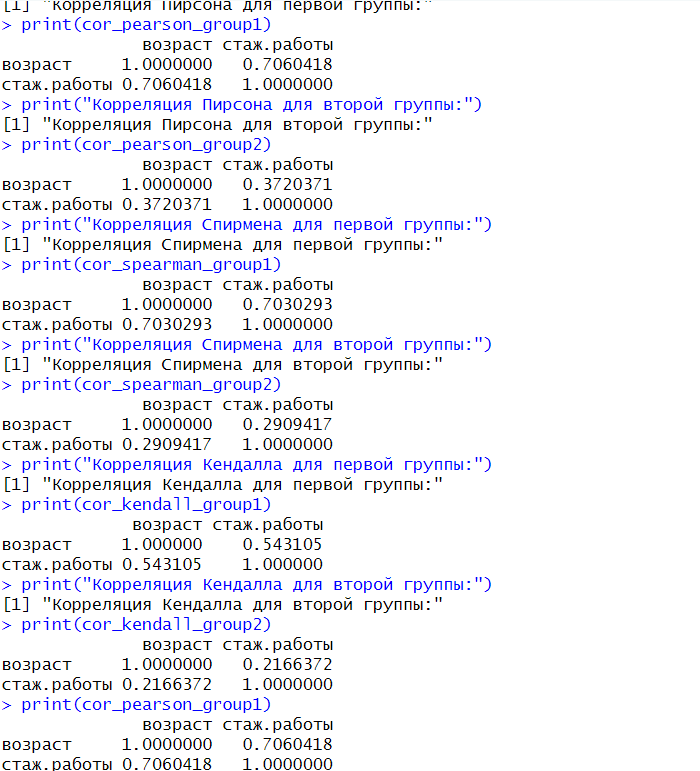
print("Корреляция Кендалла для первой группы:")

print(cor\_kendall\_group1)

print("Корреляция Кендалла для второй группы:")

print(cor\_kendall\_group2)

print(cor\_pearson\_group1)



1. **Корреляция Пирсона:**
   * **Первая группа:** Сильная положительная корреляция между возрастом и стажем работы (коэффициент = 0.706). Это указывает на значимую связь: с увеличением возраста увеличивается стаж работы.
   * **Вторая группа:** Умеренная положительная корреляция между возрастом и стажем работы (коэффициент = 0.372). Связь есть, но она менее выраженная.
2. **Корреляция Спирмена:**
   * **Первая группа:** Коэффициент 0.703 показывает сильную положительную связь между возрастом и стажем, аналогично корреляции Пирсона.
   * **Вторая группа:** Корреляция слабее (коэффициент = 0.291), но всё же положительная. Это указывает на менее выраженную зависимость.
3. **Корреляция Кендалла:**
   * **Первая группа:** Коэффициент 0.543 подтверждает значительную положительную связь между возрастом и стажем.
   * **Вторая группа:** Очень слабая положительная корреляция (коэффициент = 0.217), что показывает ещё более слабую связь по сравнению с Пирсоном и Спирменом.

**Заключение**

* В первой группе возраст и стаж работы имеют сильную взаимосвязь по всем методам корреляции.
* Во второй группе связь между возрастом и стажем слабее и менее выраженная, особенно по коэффициентам Спирмена и Кендалла.
* Корреляция Пирсона более чувствительна к линейным зависимостям, тогда как Спирмен и Кендалл подходят для ранговых зависимостей и дают более устойчивые результаты при наличии выбросов.
  1. Оценим степень взаимосвязи между двумя количественными переменными (для которых коэффициент корреляции Пирсона максимален по модулю) на основе расчета частного коэффициента корреляции для первой и второй групп.

Третью переменную выберем «стаж».

Выведем матрицу корреляции для первой группы.

print(cor\_pearson\_group1)

Чтобы найти максимальный по модулю коэффициент корреляции, нужно исключить диагональные элементы (они равны 1, что означает полную корреляцию переменной с самой собой) и найти наибольшее абсолютное значение среди оставшихся. Рассчитать частный коэффициент корреляции между переменными возраст и стаж.работы, контролируя влияние переменной ср.кол.страниц

Рассчитаем частный коэффициент корреляции через библиотеку ppcor:

*Для второй группы:*

print(cor\_pearson\_group2)

Рассчитаем частный коэффициент корреляции через библиотеку ppcor:

* 1. Графически представим матрицы коэффициентов корреляции:

**Код:**

# Выбор количественных переменных для первой группы

quantitative\_vars\_group1 <- table1[, c("возраст",

"стаж",

"кол.страниц",

"доход",

"активность.балл.оц")]

# Выбор количественных переменных для второй группы

quantitative\_vars\_group2 <- table2[, c("возраст",

"стаж",

"кол.страниц",

"доход",

"активность.балл.оц")]

# Рассчитываем коэффициенты корреляции Пирсона

cor\_pearson\_group1 <- cor(quantitative\_vars\_group1, method = "pearson")

cor\_pearson\_group2 <- cor(quantitative\_vars\_group2, method = "pearson")

# Вывод результатов корреляций

print("Корреляция Пирсона для первой группы:")

print(cor\_pearson\_group1)

print("Корреляция Пирсона для второй группы:")

print(cor\_pearson\_group2)

# Построение матричного графика для первой группы

ggpairs(quantitative\_vars\_group1,

title = "Матрица корреляции для первой группы",

lower = list(continuous = wrap("points", alpha = 0.3)),

upper = list(continuous = wrap("cor", size = 5, color = "#BB6")))

# Построение матричного графика для второй группы

ggpairs(quantitative\_vars\_group2,

title = "Матрица корреляции для второй группы",

lower = list(continuous = wrap("points", alpha = 0.3)),

upper = list(continuous = wrap("cor", size = 5, color = "#44A")))

# Функция для получения p-значений корреляции

get\_p\_value <- function(x, y) {

test <- cor.test(x, y)

return(test$p.value)

}

# Создание матрицы p-значений для первой группы

p\_matrix\_group1 <- outer(quantitative\_vars\_group1, quantitative\_vars\_group1, Vectorize(get\_p\_value))

# Создание матрицы p-значений для второй группы

p\_matrix\_group2 <- outer(quantitative\_vars\_group2, quantitative\_vars\_group2, Vectorize(get\_p\_value))

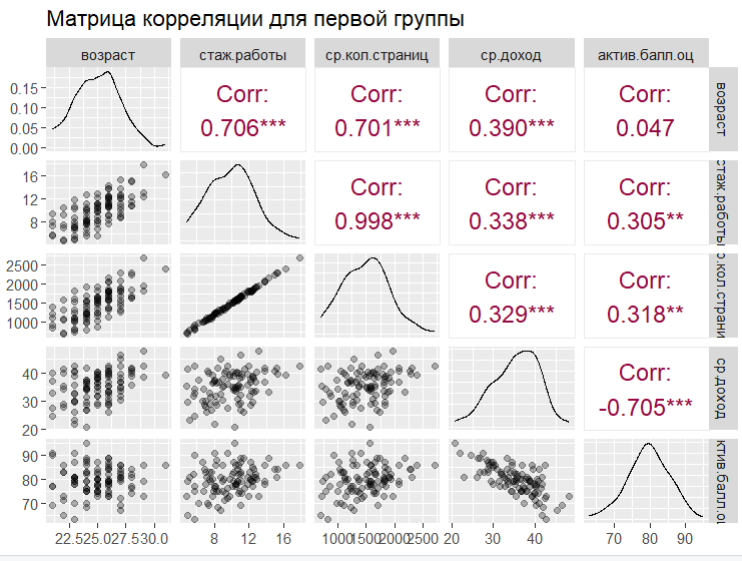
# Вывод p-значений

print("P-значения для первой группы:")

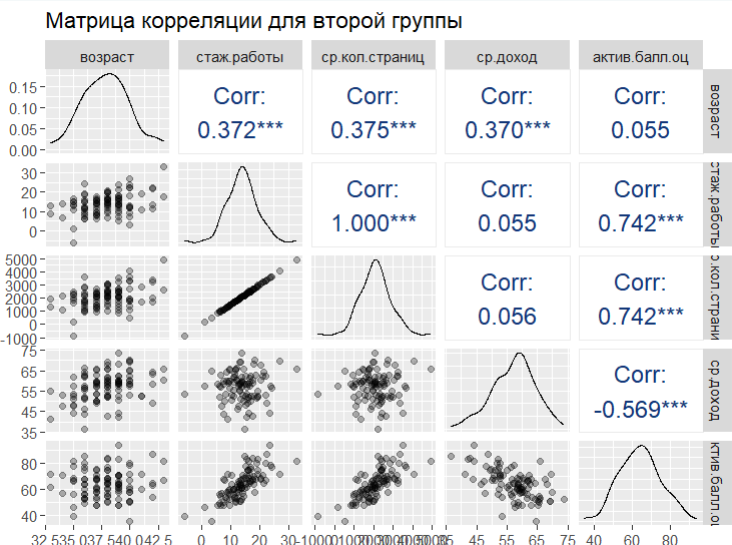
print(p\_matrix\_group1)

print("P-значения для второй группы:")

print(p\_matrix\_group2)



*Рис. 12 Матрица коэффициентов корреляции для первой группы*

**

*Рис. 13 Матрица коэффициентов корреляции для второй группы*

1. **Возраст и стаж работы**:
   * В первой группе корреляция между возрастом и стажем работы довольно высокая (0.706), что естественно, так как с возрастом стаж обычно увеличивается.
   * Во второй группе эта корреляция также положительная, но значительно ниже (0.372), что может указывать на большую изменчивость стажа работы относительно возраста в этой группе.
2. **Возраст и среднее количество страниц**:
   * В первой группе корреляция между возрастом и средним количеством страниц довольно высокая (0.701), тогда как во второй группе эта связь слабее (0.375).
   * Это может означать, что в первой группе с возрастом в среднем увеличивается количество страниц (возможно, документов или отчетов), тогда как во второй группе эта связь выражена слабее.
3. **Средний доход и активы**:
   * В первой группе имеется сильная отрицательная корреляция между средним доходом и активами (-0.705), что может свидетельствовать о том, что при высоком доходе уровень активов относительно ниже.
   * Во второй группе аналогичная отрицательная корреляция, но чуть слабее (-0.569). Это может указывать на схожую тенденцию, но менее выраженную.
4. **Стаж работы и среднее количество страниц**:
   * В обеих группах корреляция между стажем работы и средним количеством страниц очень высокая (практически 1), что говорит о практически линейной связи между этими переменными.
5. **Средний доход и стаж работы**:
   * В первой группе имеется умеренная положительная корреляция между стажем работы и средним доходом (0.338), что говорит о том, что доход возрастает с увеличением стажа.
   * Во второй группе эта корреляция гораздо слабее (0.055), что может означать меньшую зависимость дохода от стажа работы.
6. **Среднее количество страниц и активы**:
   * Во второй группе между средним количеством страниц и активами наблюдается умеренная положительная корреляция (0.742), в то время как в первой группе эта связь слабее (0.318).
   * Это может свидетельствовать о том, что вторая группа больше ассоциирует количество страниц с активами.

**Выводы**:

* В первой группе прослеживается более сильная связь между возрастом, стажем и количеством страниц, а также отрицательная связь между средним доходом и активами.
* Во второй группе наблюдаются более слабые связи между возрастом и другими показателями, но усиленная связь между количеством страниц и активами.
* Возможно, группы отличаются по характеру работы или социальной структуре, что и вызывает различия в этих взаимосвязях.

1. Выводы

В ходе выполнения лабораторной работы был проведен всесторонний анализ данных, относящихся к профессиональной деятельности двух групп сотрудников. Ключевым этапом стало изучение статистических характеристик, таких как минимальные и максимальные значения, среднее, стандартное отклонение, квартили и медиана. Эти показатели позволили лучше понять особенности распределения данных как внутри каждой группы, так и в общем по выборке.

Графическое представление результатов с помощью различных диаграмм и гистограмм обеспечило наглядную иллюстрацию распределений и помогло выявить возможные паттерны. Диаграммы размаха дополнительно предоставили информацию о вариациях в данных по отдельным категориям.

Корреляционный анализ позволил определить степень взаимосвязи между несколькими переменными. Были рассчитаны коэффициенты корреляции Пирсона, Спирмена и Кендалла, что дало возможность оценить как силу, так и направление этих связей.