МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра вычислительной техники

**Лабораторная работа №1**

**по дисциплине «Методы анализа данных»**

**ВВЕДЕНИЕ В R. МЕТОДЫ ПЕРВИЧНОГО РАЗВЕДОЧНОГО АНАЛИЗА ДАННЫХ В R**

|  |  |
| --- | --- |
| Студент | Великжанин С.И. |
| Вариант | 3 |
| Группа | АВТ-010 |
| Преподаватель | Альсова О.К. |
| Дата сдачи | 30.10.2023 |

Новосибирск, 2023 г.

1. **Цели**

- 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).

1. **Постановка задачи ИАД**

Изучаются показатели работы программистов крупной организации.Рассматриваются следующие показатели (признаки) для каждого программиста:

* пол (1-м, 2-ж);
* возраст;
* стаж работы;
* процент разработок, выполненных в срок в рамках бюджета с требуемым функционалом (за год);
* количество ошибок, выявленных пользователем (за год);
* стаж работы по специальности в данной организации;
* степень удовлетворенности заказчика;
* качество документирования (1 – низкое, 2 – среднее, 3 – выше среднего, 4- высокое).

Необходимо провести предварительный разведочный анализ данных с целью описания характера распределения данных, выявления структуры взаимосвязей между показателями.

Программисты разбиты на две группы в зависимости от стажа работы.

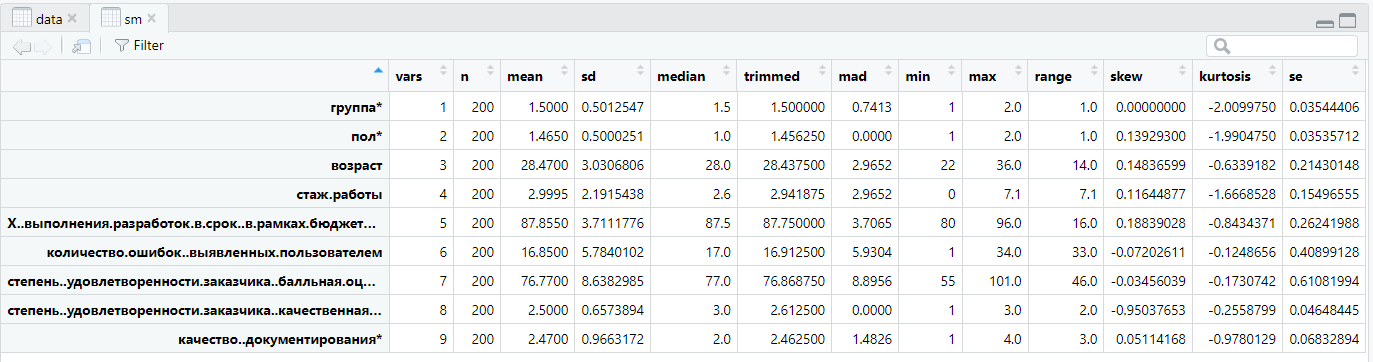
**Вариант 3**

|  |
| --- |
| 1 группа - стаж менее 2,5 |
| 2 группа - стаж более 2,5 |

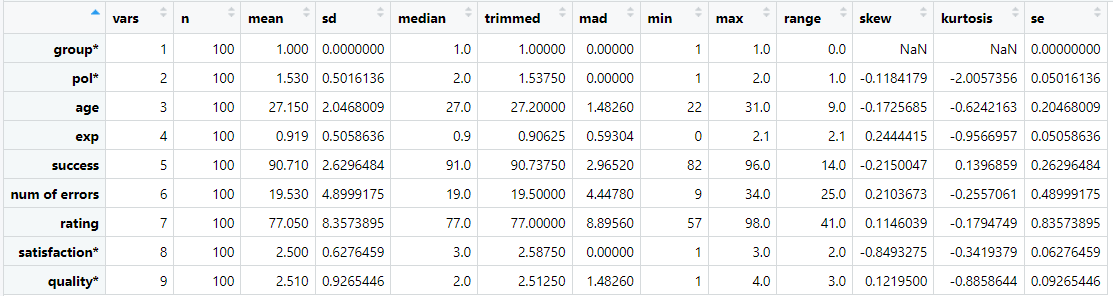
1. **Ход работы**

Расчет основных статистических характеристик:

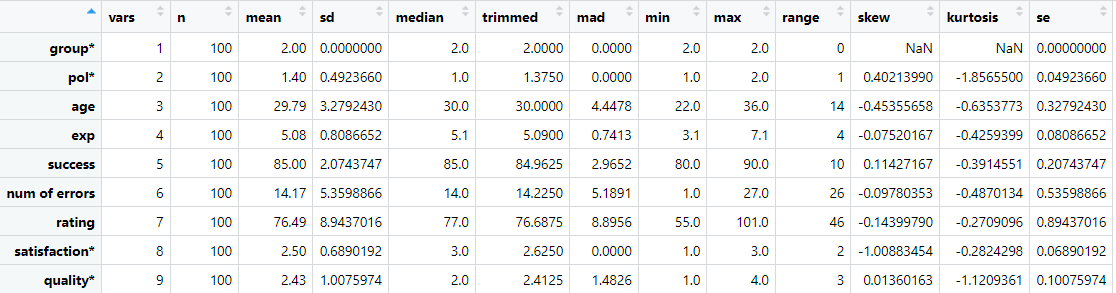
- Для всех программистов



- Для первой группы



- Для второй группы

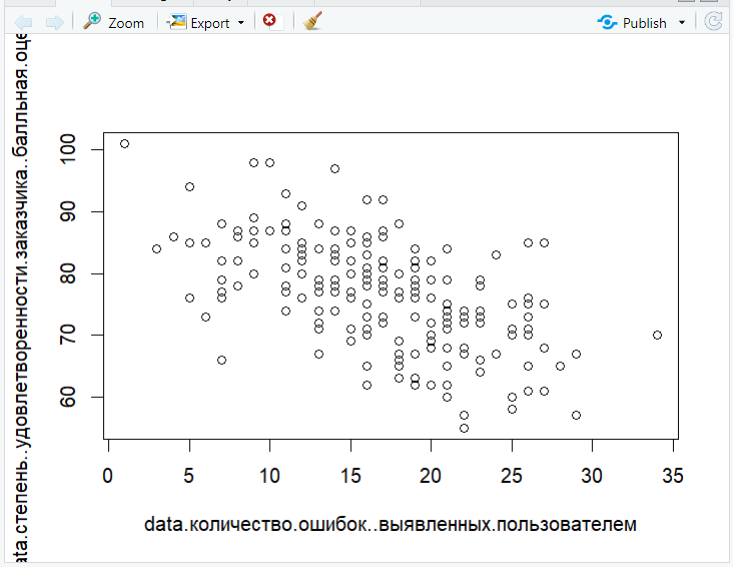


Вывод: Данные показывают, что у группы с меньшим средним опытом работы среди сотрудников выше частота выполнения работ в срок, однако при этом они допускают больше ошибок, чем сотрудники с опытом работы более 2,5 лет.

**Графический анализ данных:**

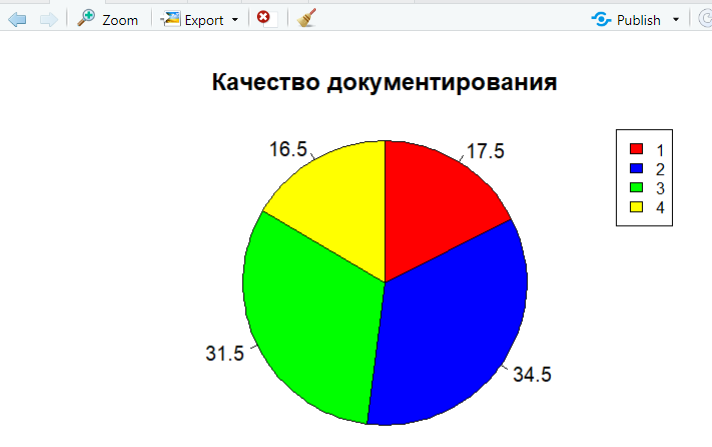
Диаграмма рассеяния бальной оценки от кол-ва ошибок

> plot(data.frame(data$количество.ошибок..выявленных.пользователем,data$степень..удовлетворенности.заказчика..балльная.оценка.))



Радиальная диаграмма качества документирования (от 1 до 4):

> pie(x, piepercent, radius=1, main="Качество документирования", col=c("red", "blue", "green", "yellow"), clockwise=TRUE)

> legend("topright", c("1","2","3","4"), cex = 0.8, fill =c("red", "blue", "green", "yellow"))

Качество документирования по группам и полу (против часовой стрелки):

> x1<-c(summary(data$качество..документирования[data$группа=="1"&data$пол=="1"]))

> x2<-c(summary(data$качество..документирования[data$группа=="1"&data$пол=="2"]))

> x3<-c(summary(data$качество..документирования[data$группа=="2"&data$пол=="1"]))

> x4<-c(summary(data$качество..документирования[data$группа=="2"&data$пол=="2"]))

> piepercent1<-round(100\*x1/200,1)

> piepercent2<-round(100\*x2/200,1)

> piepercent3<-round(100\*x3/200,1)

> piepercent4<-round(100\*x4/200,1)

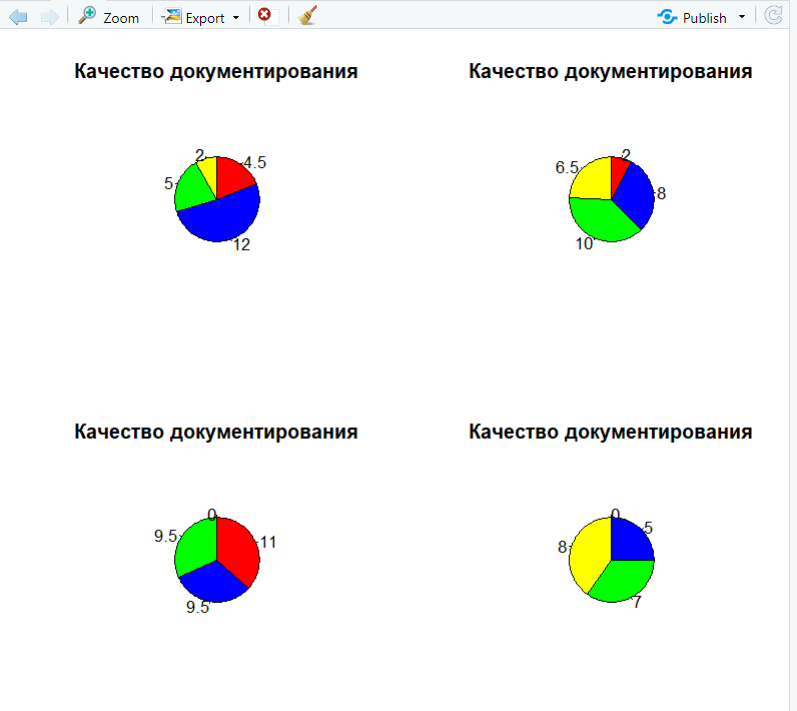
> par(mfrow=c(2,2))

> pie(x1, piepercent1, radius=1, main="Качество документирования", col=c("red", "blue", "green", "yellow"), clockwise=TRUE)

> pie(x2, piepercent2, radius=1, main="Качество документирования", col=c("red", "blue", "green", "yellow"), clockwise=TRUE)

> pie(x3, piepercent3, radius=1, main="Качество документирования", col=c("red", "blue", "green", "yellow"), clockwise=TRUE)

> pie(x4, piepercent4, radius=1, main="Качество документирования", col=c("red", "blue", "green", "yellow"), clockwise=TRUE)



Столбиковая диаграмма возраста по группам и полу:

> barplot(table(data$возраст[data$группа=="1"&data$пол=="1"]), col=c("red"), main="1гр ж",xlab="Возраст")

> barplot(table(data$возраст[data$группа=="1"&data$пол=="2"]), col=c("red"), main="1гр м",xlab="Возраст")

> barplot(table(data$возраст[data$группа=="2"&data$пол=="1"]), col=c("red"), main="2гр ж",xlab="Возраст")

> barplot(table(data$возраст[data$группа=="2"&data$пол=="2"]), col=c("red"), main="2гр м",xlab="Возраст")

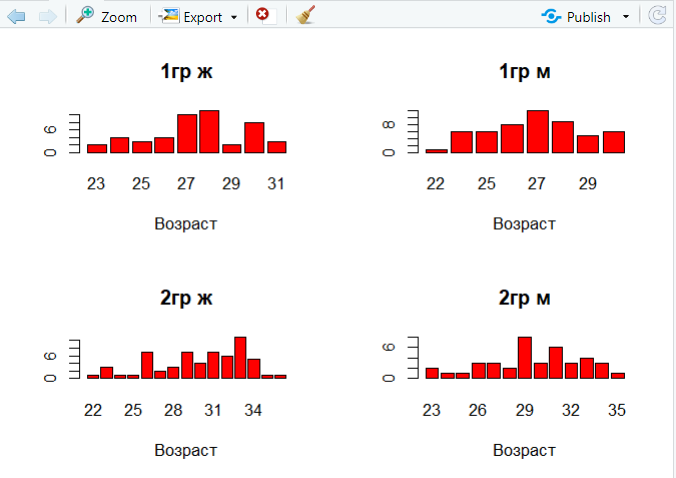
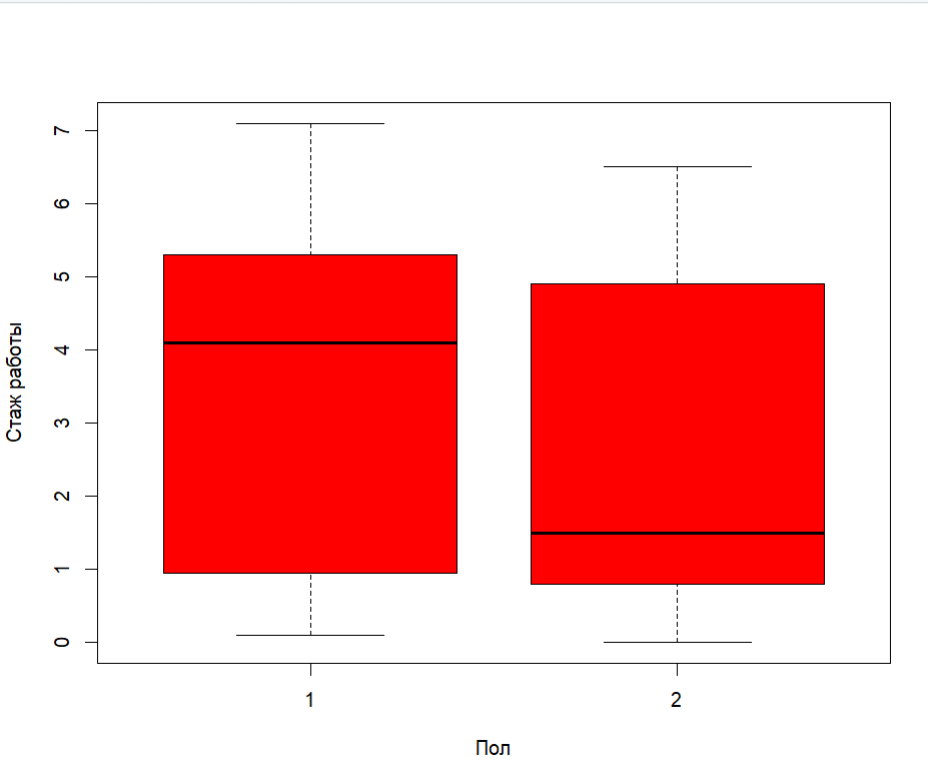


Диаграмма размаха пол - стаж работы:

> boxplot(data$стаж.работы ~ data$пол, xlab="Пол", ylab="Стаж работы",col="red", data=data)



Гистограммы количественных признаков:

> hist(data$возраст, freq=FALSE, breaks=12, col="red", xlab="Возраст", main="Гистограмма")

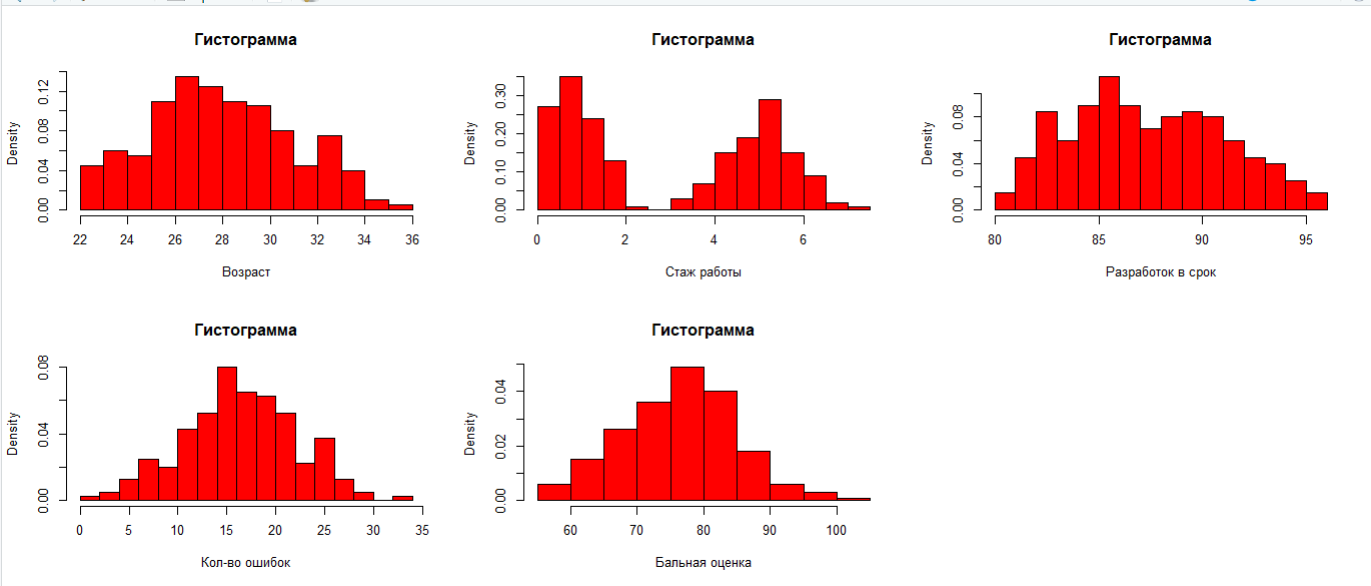
> hist(data$стаж.работы, freq=FALSE, breaks=12, col="red", xlab="Стаж работы", main="Гистограмма")

> hist(data$X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом, freq=FALSE, breaks=12, col="red", xlab="Разработок в срок", main="Гистограмма")

> hist(data$количество.ошибок..выявленных.пользователем, freq=FALSE, breaks=12, col="red", xlab="Кол-во ошибок", main="Гистограмма")

> hist(data$степень..удовлетворенности.заказчика..балльная.оценка., freq=FALSE, breaks=12, col="red", xlab="Бальная оценка", main="Гистограмма")

> hist(data$качество..документирования, freq=FALSE, breaks=12, col="red", xlab="Качество документирования", main="Гистограмма")



Матричный график количественных переменных:

> pairs(~data$возраст+data$стаж.работы+data$X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом+data$количество.ошибок..выявленных.пользователем+data$степень..удовлетворенности.заказчика..балльная.оценка., data=data, main="Матричный график")



Выводы: Графический анализ показывает обратную зависимость между количеством ошибок и оценкой проделанной работы, а также факт того, что у группы с большим опытом работы выше качество документирования. В компании больше всего сотрудников в возрасте 25-29 лет.

**Корреляционный анализ данных:**

Оценка степени взаимосвязи между качественными переменными для первой группы:

> group1 <- subset(data, группа==1, select=c("пол", "качество..документирования"))

> View(group1)

> group2 <- subset(data, группа==2, select=c("пол", "качество..документирования"))

> table<-table(group1$пол, group1$качество..документирования)

> chisq.test(table(group1$пол, group1$качество..документирования))

Pearson's Chi-squared test

data: table(group1$пол, group1$качество..документирования)

**X-squared = 11.302**, df = 3, p-value = 0.0102

> fisher.test(table(group1$пол, group1$качество..документирования))

Fisher's Exact Test for Count Data

data: table(group1$пол, group1$качество..документирования)

**p-value = 0.01029**

alternative hypothesis: two.sided

Для второй группы:

> table<-table(group2$пол, group2$качество..документирования)

> chisq.test(table(group2$пол, group2$качество..документирования))

Pearson's Chi-squared test

data: table(group2$пол, group2$качество..документирования)

**X-squared = 39.115**, df = 3, p-value = 1.641e-08

> fisher.test(table(group2$пол, group2$качество..документирования))

Fisher's Exact Test for Count Data

data: table(group2$пол, group2$качество..документирования)

**p-value = 2.47e-10**

alternative hypothesis: two.sided

Однофакторный дисперсионный анализ для пола и двух количественных х-к:

> aov\_model <- aov(**количество.ошибок..выявленных.пользователем ~ пол**, data=data)

> summary(aov\_model)

Df Sum Sq Mean Sq F value Pr(>F)

пол 1 30 30.49 0.911 0.341

Residuals 198 6627 33.47

> aov\_model <- aov(**степень..удовлетворенности.заказчика..балльная.оценка. ~ пол**, data=data)

> summary(aov\_model)

Df Sum Sq Mean Sq F value Pr(>F)

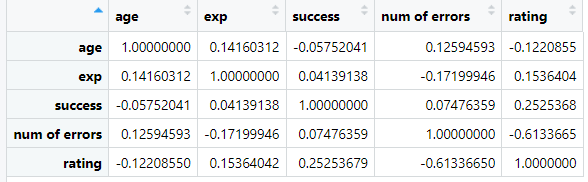
пол 1 407 407.5 5.587 0.0191 \*

Residuals 198 14442 72.9

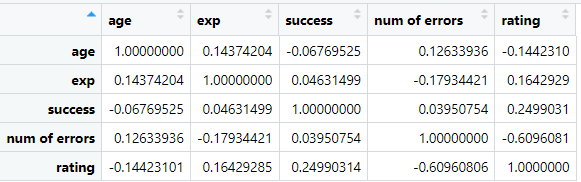
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

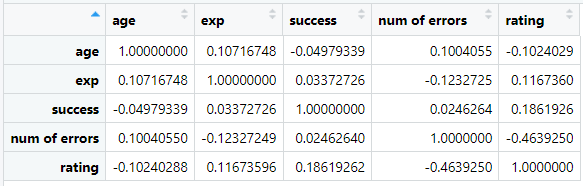
Корреляция по Пирсону для 1 группы:



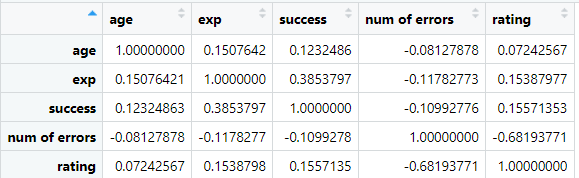
Корреляция по Спирмену для 1 группы:



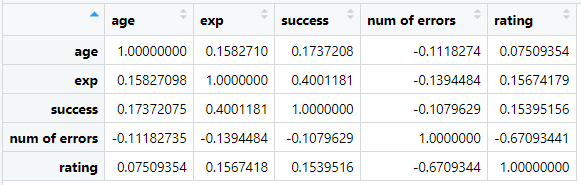
Корреляция по Кендаллу для 1 группы:



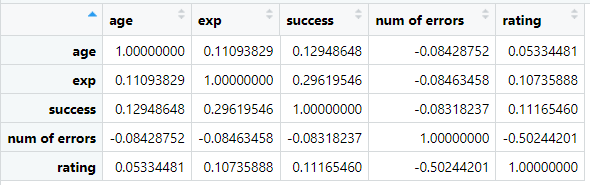
Корреляция по Пирсону для 2 группы:



Корреляция по Спирмену для 2 группы:



Корреляция по Кендаллу для 2 группы:



Оценим степень взаимосвязи между количеством ошибок и бальной оценкой на основе частного коэф. Корреляции:

> M<-**group1**[,unlist(lapply(data, is.numeric))]

> library(ggm)

> pcor(c(4,5,1,2,3), cov(M))

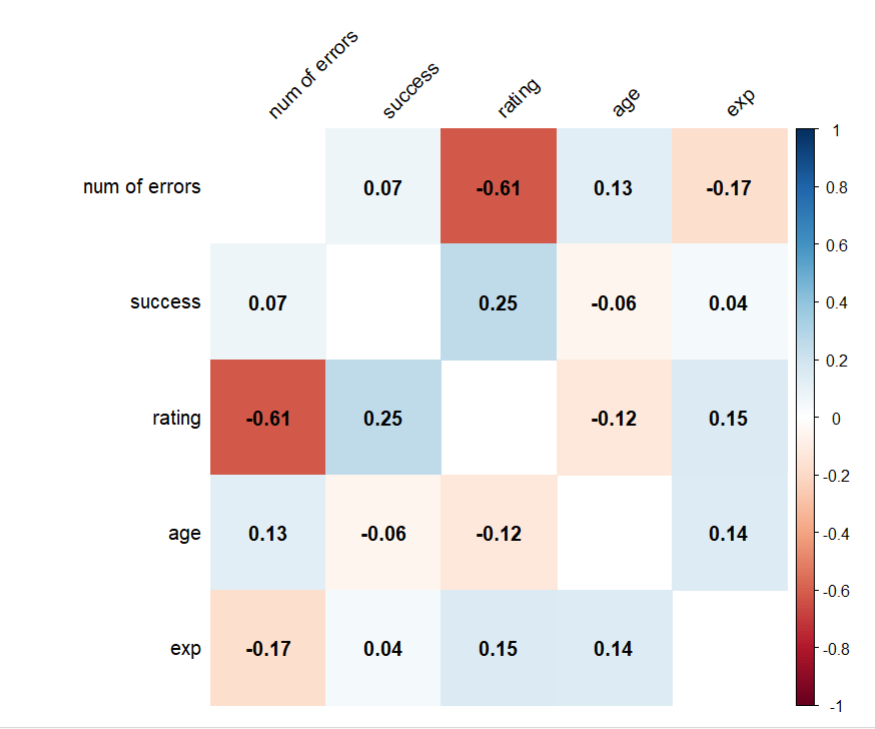
**[1] -0.6386486**

> M<-**group2**[,unlist(lapply(data, is.numeric))]

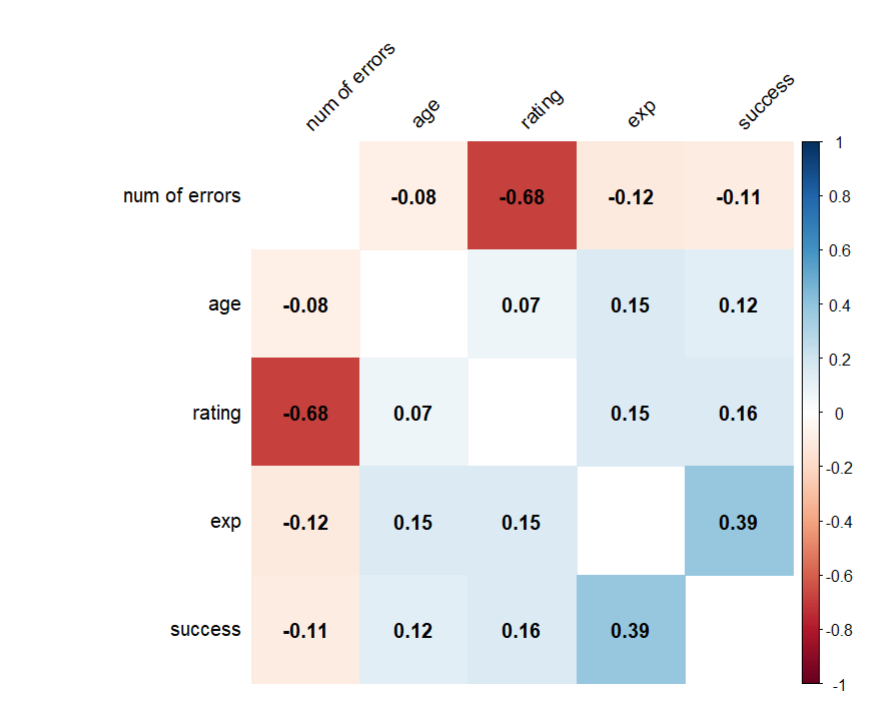
> pcor(c(4,5,1,2,3), cov(M))

**[1] -0.6737715**

Матрица коэффициентов корреляции для первой группы:



Матрица коэффициентов корреляции для второй группы:



Статистическая значимость связи:

> cor.test(N11, N21, alternative = **"two.side")**

Pearson's product-moment correlation

data: N11 and N21

t = 14.271, df = 23, p-value = 6.465e-13

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.8838027 0.9770767

sample estimates:

**cor**

**0.9479075**

> cor.test(N11, N21, alternative = "**less**")

Pearson's product-moment correlation

data: N11 and N21

t = 14.271, df = 23, p-value = 1

alternative hypothesis: true correlation is less than 0

95 percent confidence interval:

-1.0000000 0.9738233

sample estimates:

**cor**

**0.9479075**

> cor.test(N11, N21, alternative = "**greater**")

Pearson's product-moment correlation

data: N11 and N21

t = 14.271, df = 23, p-value = 3.233e-13

alternative hypothesis: true correlation is greater than 0

95 percent confidence interval:

0.8976644 1.0000000

sample estimates:

**cor**

**0.9479075**

1. **Вывод**

В ходе выполнения работы были получены навыки работы в среде R, изучены средства для проведения первичного разведочного анализа данных.

* Были изучены показатели работы программистов в большой компании. Выборка состояла из 200 сотрудников, разделенных на 2 группы по стажу работы в компании (больше и меньше 2,5 лет).
* Выявлена прямая связь между опытом работы и количеством допущенных ошибок при выполнении задания, качеством документирования.
* Выявлена обратная связь между опытом работы и скоростью выполнения задания (чем меньше опыт – тем меньше заняло времени).
* Количество успешно выполненных заданий сотрудника напрямую зависит от опыта работы в компании.