**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ**

**РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ**

**ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

**КАФЕДРА ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ**

Лабораторная работа №1 «

Введение в R. Методы первичного разведочного анализа данных в R»

по дисциплине

«Анализ и интерпретация данных»

Группа: АПИМ-24

Факультет: АВТФ

Студенты: Разуваев В.В.б Чернощеков В.---

Преподаватель: Альсова О.К.

Вариант: 4

Содержание

[Цель работы 3](#_Toc147682089)

[Задание 3](#_Toc147682090)

[Расчет основных статистических характеристик для первой и второй группы и для всей выборки. Выводы. 5](#_Toc147682091)

[Графический разведочный анализ данных 8](#_Toc147682092)

[Корреляционный анализ данных 14](#_Toc147682093)

[Вывод 19](#_Toc147682094)

[Приложение 20](#_Toc147682095)

## Цель работы

* получить базовые навыки работы в среде R;
* получить навыки применения методов первичного разведочного ана-  
  лиза данных и статистической проверки гипотез для решения задач   
  АД;
* 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).

## Задание

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

1. Установите рабочую директорию. 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 и 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. Проведите корреляционный анализ данных. Проверьте гипотезу о соответствии данных модели нормального за-  
   кона распределения для одного из количественных признаков (от-  
   дельно 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).

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.

7. Сделайте выводы по работе в произвольной форме.

**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 4

**1 группа – стаж менее 5**

**2 группа - стаж более 5**

## Расчет основных статистических характеристик для первой и второй группы и для всей выборки. Выводы.

Основные статистические характеристики по количественным данным (минимальное, максимальное, среднее значение, стандартное отклонение, первый и третий квартили, медиана, мода, асимметрия, эксцесс) для всей выборки, для первой группы, для второй группы представлены в таблицах 1, 2, 3 соответственно.

Таблица 1. Основные статистические характеристики для всей выборки

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Мин. | Макс. | Среднее | Стандартное отклонение | Q1 | Q3 | Медиана | Мода | Ассиметрия | Эксцесс |
| Возраст | 22.0 | 37.0 | 29.51 | 2.9484516 | 27.0 | 31.0 | 30.0 | 30 | 0.02508594 | -0.5035062 |
| Стаж работы | 0.8 | 8.8 | 3.911 | 2.1098665 | 1.9 | 5.8 | 3.6 | 2.2, 2 | 0.20578811 | -1.4967055 |
| % выполнения разработок в срок, в рамках бюджета, с требуемым функционалом | 73.0 | 100.0 | 88.985 | 5.3335739 | 85.0 | 93.0 | 89.0 | 92, 88 | -0.22423104 | -0.3712056 |
| Количество ошибок, выявленных пользователем | 1.0 | 29.0 | 15.885 | 5.4044254 | 12.0 | 19.25 | 16.0 | 17 | -0.0962751 | -0.1938481 |
| Степень удовлетворенности заказчика (балльная оценка) | 55.0 | 99.0 | 78.315 | 10.2342304 | 70.0 | 86.25 | 79.0 | 69 | -0.04448076 | -0.8574756 |
| Качество документирования | 1.0 | 4.0 | 2.255 | 0.9077970 | 2.0 | 3.0 | 2.0 | 3 | 0.04162181 | -0.9840513 |

Таблица 2. Основные статистические характеристики для первой группы

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Мин. | Макс. | Среднее | Стандартное отклонение | Q1 | Q3 | Медиана | Мода | Ассиметрия | Эксцесс |
| Возраст | 22.0 | 32.0 | 27.470 | 2.0521238 | 26.0 | 29.0 | 27.0 | 27 | -0.19350472 | -0.2640987 |
| Стаж работы | 0.8 | 3.5 | 1.936 | 0.4814792 | 1.6 | 2.2 | 1.9 | 2.2, 2 | 0.51503457 | 0.5489041 |
| % выполнения разработок в срок, в рамках бюджета, с требуемым функционалом | 73.0 | 98.0 | 86.740 | 5.1455973 | 83.0 | 90.0 | 87.0 | 89, 85, 86 | -0.13488429 | -0.4084072 |
| Количество ошибок, выявленных пользователем | 6.0 | 29.0 | 17.540 | 4.9347664 | 14.0 | 21.0 | 17.0 | 17 | -0.04920547 | -0.3986077 |
| Степень удовлетворенности заказчика (балльная оценка) | 56.0 | 99.0 | 74.100 | 8.8346197 | 67.0 | 80.25 | 73.5 | 78, 66 | 0.23931380 | -0.5931316 |
| Качество документирования | 1.0 | 4.0 | 2.190 | 0.8491826 | 2.0 | 3.0 | 2.0 | 2 | 0.02648464 | -0.9581964 |

Таблица 3. Основные статистические характеристики для второй группы

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Мин. | Макс. | Среднее | Стандартное отклонение | Q1 | Q3 | Медиана | Мода | Ассиметрия | Эксцесс |
| Возраст | 26.0 | 37.0 | 31.550 | 2.2036471 | 30.0 | 33.0 | 31.0 | 33 | -0.088659574 | -0.2800454 |
| Стаж работы | 3.7 | 8.8 | 5.886 | 0.9144209 | 5.3 | 6.5 | 5.8 | 5.6, 6 | 0.344190387 | 0.7026029 |
| % выполнения разработок в срок, в рамках бюджета, с требуемым функционалом | 81.0 | 100.0 | 91.230 | 4.5301571 | 88.0 | 94.25 | 92.0 | 93 | -0.107142474 | -0.8527156 |
| Количество ошибок, выявленных пользователем | 1.0 | 29.0 | 14.230 | 5.3689248 | 11.0 | 17.25 | 14.0 | 17 | -0.016107181 | -0.1264653 |
| Степень удовлетворенности заказчика (балльная оценка) | 55.0 | 99.0 | 82.530 | 9.8282781 | 75.75 | 89.0 | 84.0 | 89 | -0.518799591 | -0.3881069 |
| Качество документирования | 1.0 | 4.0 | 2.320 | 0.9627402 | 1.75 | 3.0 | 2.0 | 3 | 0.007100482 | -1.0937513 |

**Выводы на основе полученных характеристик**

Минимальные возраст разработчика во второй группе (26) выше чем в первой группе (2). Также различается диапазоны стажа работы в этих группах, они не пересекаются. Признаки «% выполнения разработок в срок, в рамках бюджета, с требуемым функционалом», «Количество ошибок, выявленных пользователем», «Степень удовлетворенности заказчика (балльная оценка)» ближе к оптимуму во второй группе. Что касается качества документирования для каждой из трех выборок, то минимальное, максимальное, медианы, первая и третья квартили совпадают. Значение ассиметрии признака «Степень удовлетворенности заказчика (балльная оценка)» во второй группе отрицательное, а в первой положительное, что говорит о близости степени удовлетворенности к верхней границе во второй группе и близости степени удовлетворенности к нижней границе в первой группе.

На основе полученных результатов был проведен графический анализ. Диаграммы представлены в следующем разделе.

## Графический разведочный анализ данных

Построим диаграмму рассеяния (рис. 1) для признаков «возраст» и «стаж работы». Наблюдается прямая зависимость между возрастом и стажем.

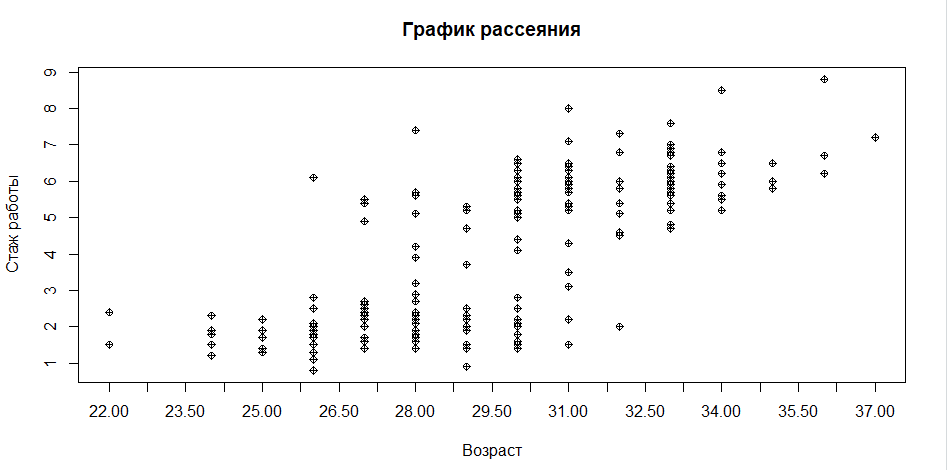


Рисунок – Диаграмма рассеяния для возраста и стажа работы

Построим радиальные диаграммы для признака ««степень удовлетворенности заказчика» (балльная оценка)» (рис. 2).

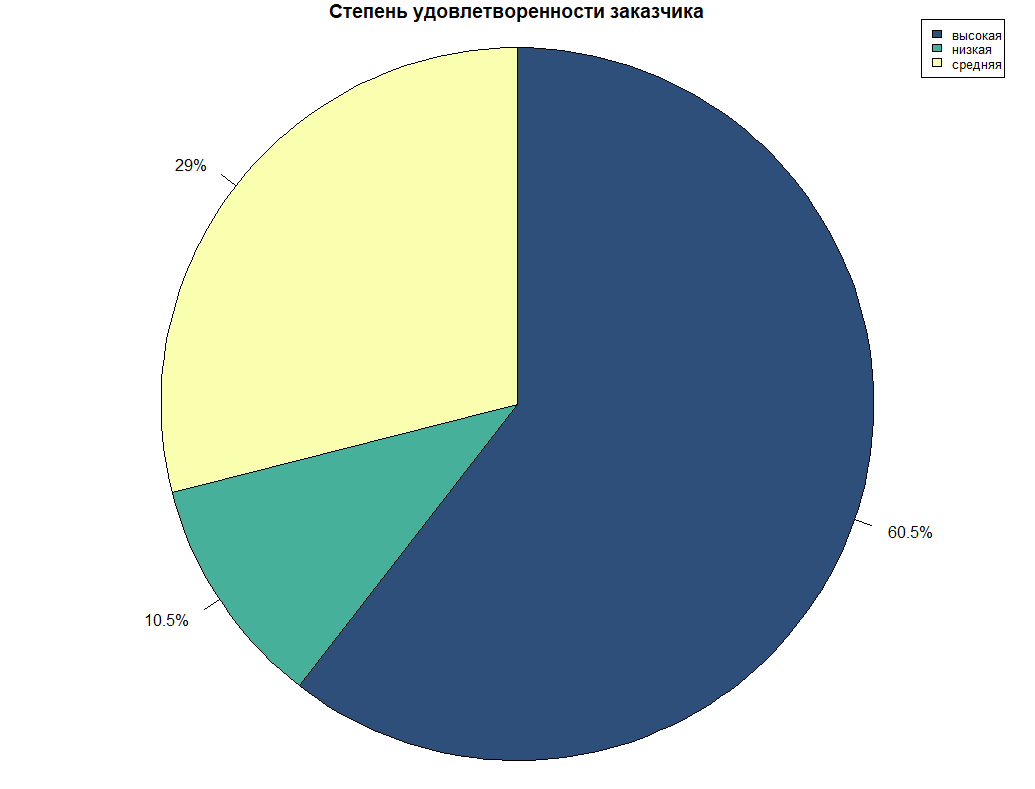


Рисунок - Радиальная диаграмма по качественному признаку «степень удовлетворенности заказчика»

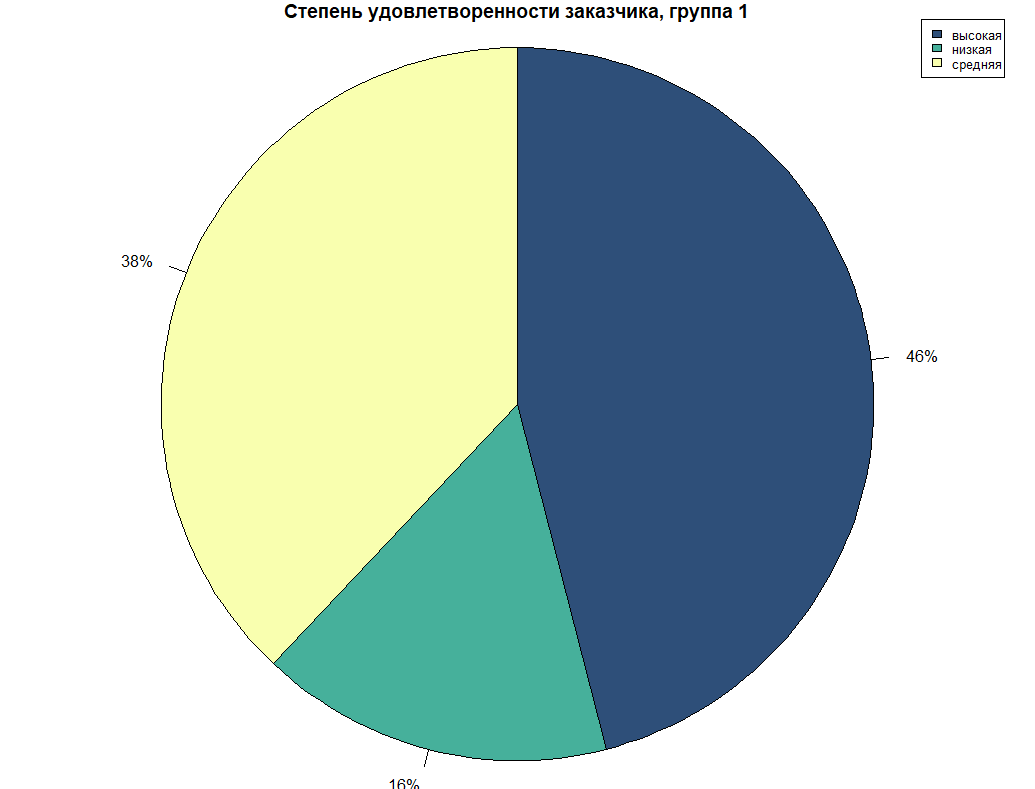


Рисунок - Радиальная диаграмма по качественному признаку «степень удовлетворенности заказчика» для группы 1

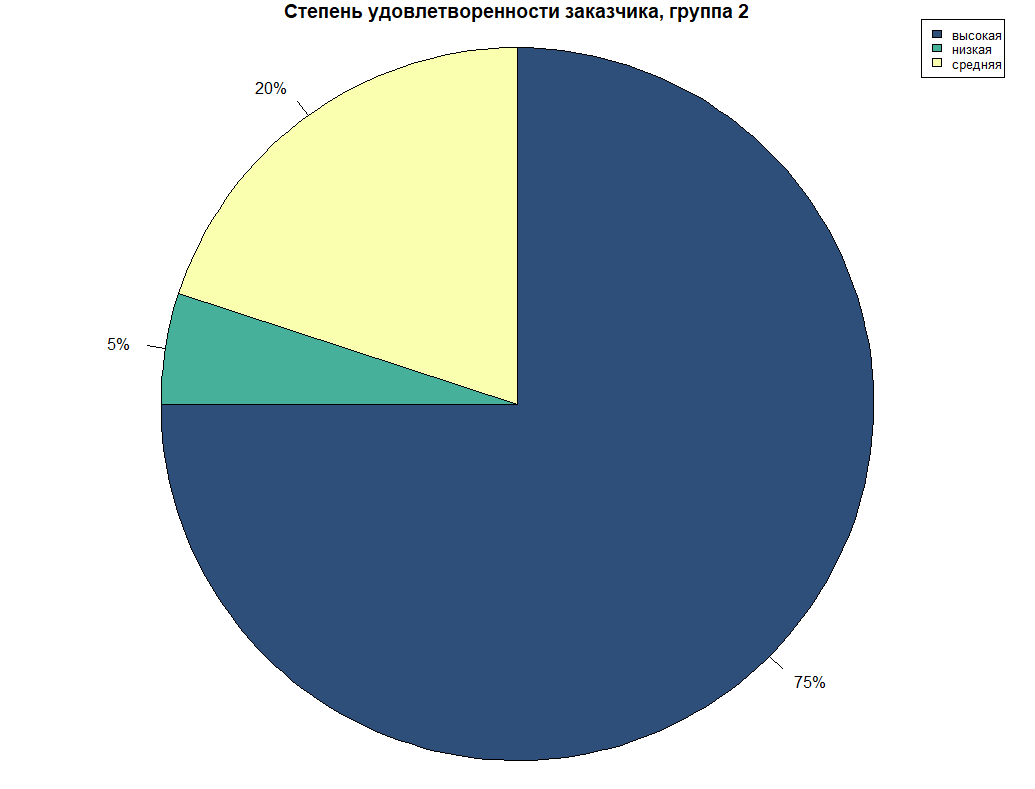


Рисунок - Радиальная диаграмма по качественному признаку «степень удовлетворенности заказчика» для группы 2

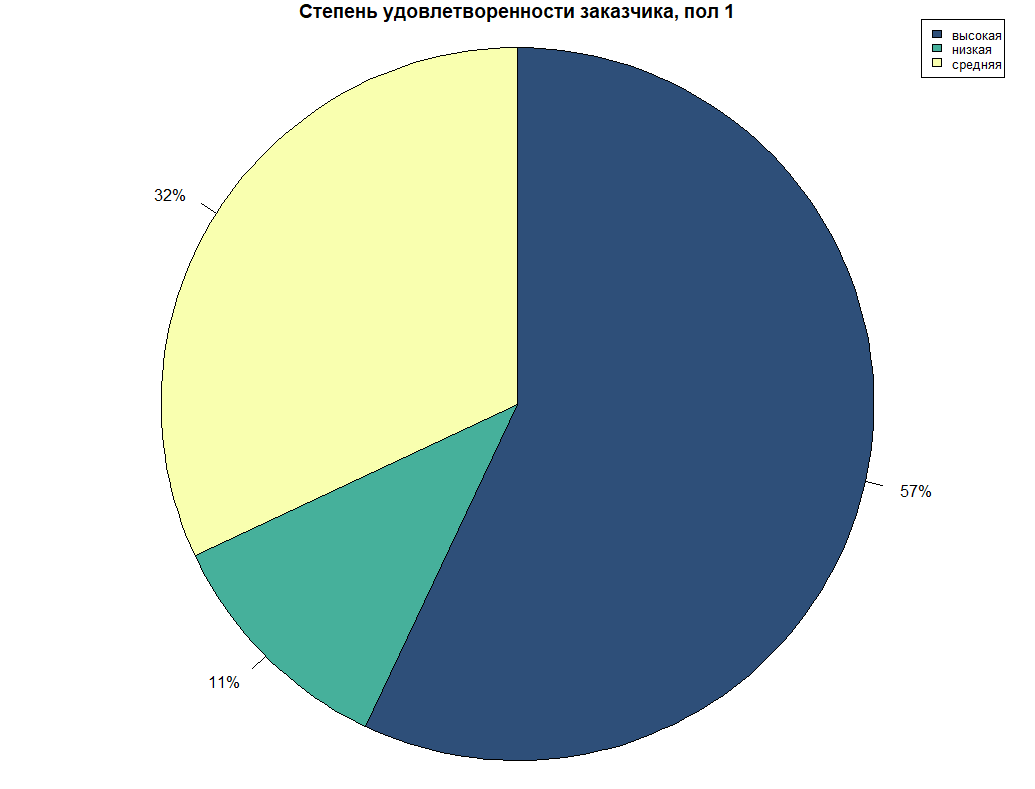


Рисунок - Радиальная диаграмма по качественному признаку «степень удовлетворенности заказчика» для пола 1

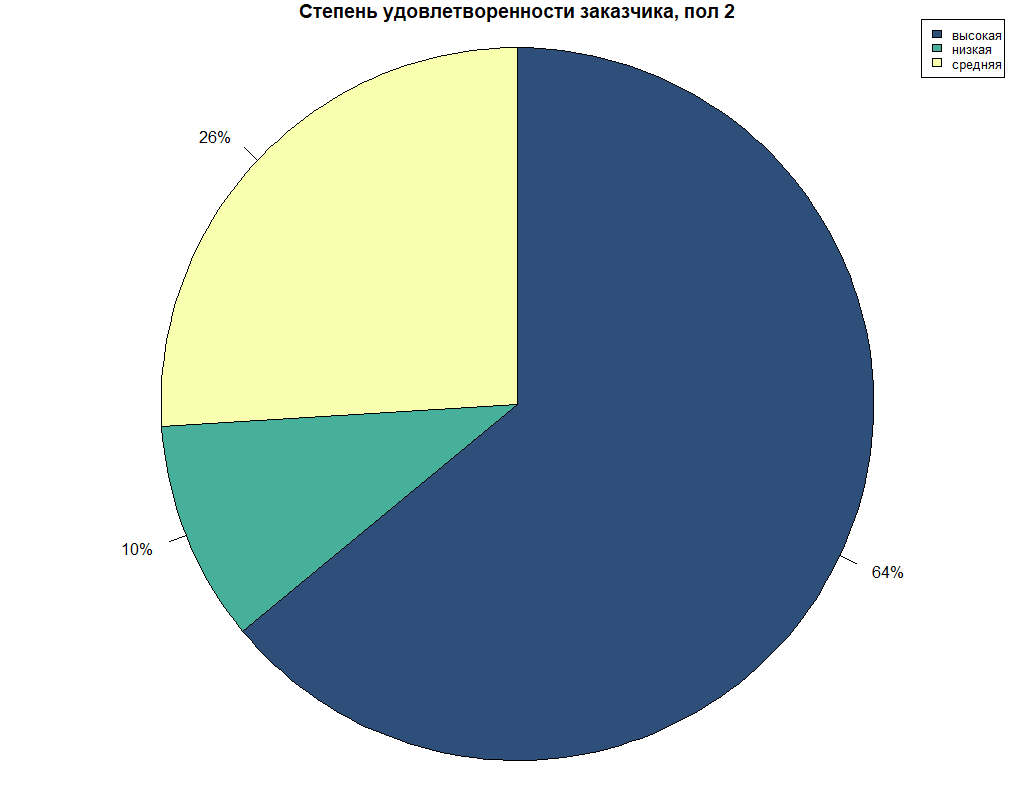


Рисунок - Радиальная диаграмма по качественному признаку «степень удовлетворенности заказчика» для пола 2

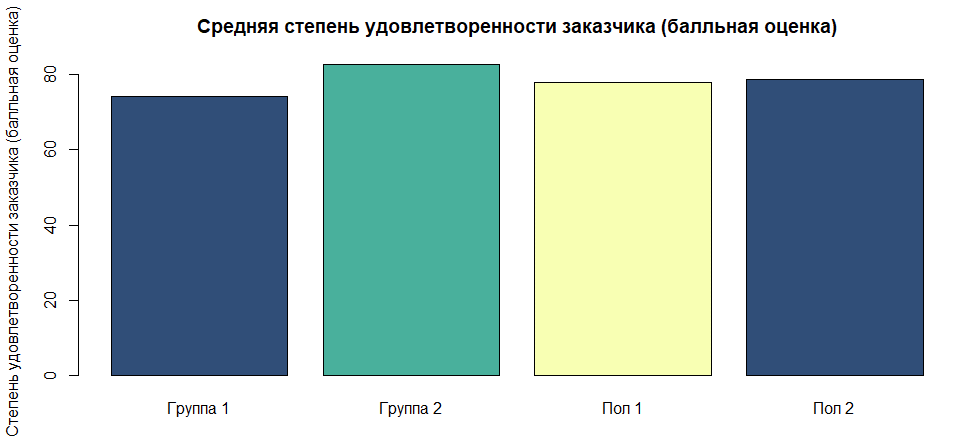


Рисунок – Столбиковая диаграмма по количественному признаку «степень удовлетворенности заказчика»

Во второй группе оценка высокая встречается в 75% случаев и значительно превышает этот показатель в первой группе (46%). В выборке с полом 2 высокая оценка встречается на 7% чаще чем в выборке с полом 1.

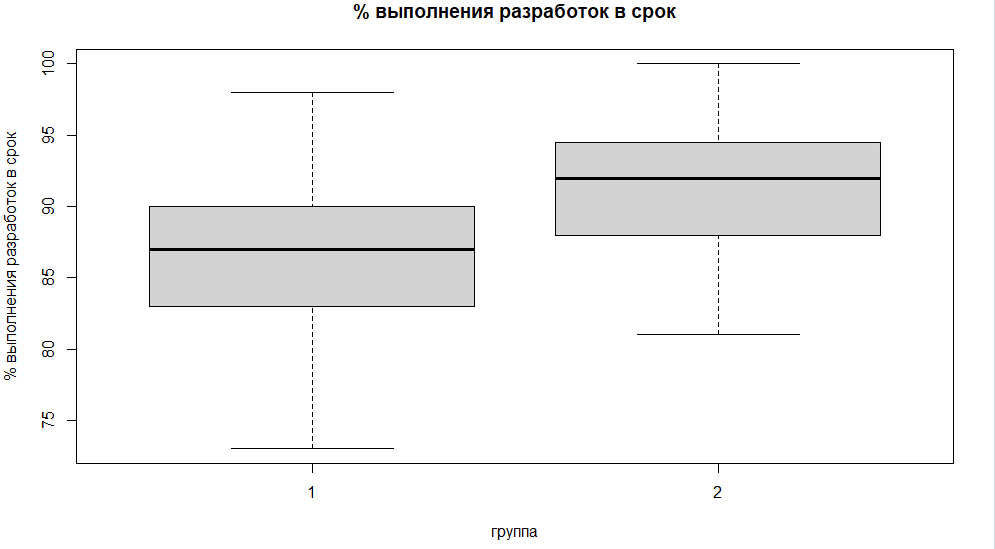


Рисунок – Диаграмма размаха по количественному признаку «% выполнения разработок в срок c требуемым функционалом» в зависимости от группы

По диаграмме размаха на рисунке 8 видно, что разработчики из второй группы выполняются в срок больше 80% работы. В среднем они успевают выполнить больше работы в срок чем разработчики из первой группы.

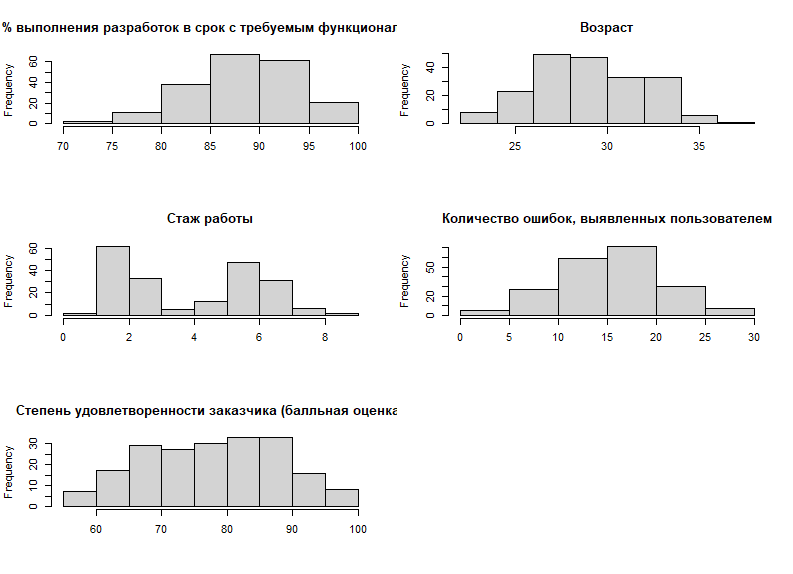


Рисунок – Гистограммы для всех количественных признаков

На рисунке 9 приведены гистограммы распределения для всех количественных признаков. В гистограммах, демонстрирующих возраст, удовлетворенность заказчика, % выполнения работ в срок, количество ошибок, выявленных пользователем распределение близко к нормальному. В гистограмме, демонстрирующей стаж работы, наблюдается разделение на две группы.

На рисунке 10 изображен матричный график, содержащий диаграммы рассеяния между количественными признаками.

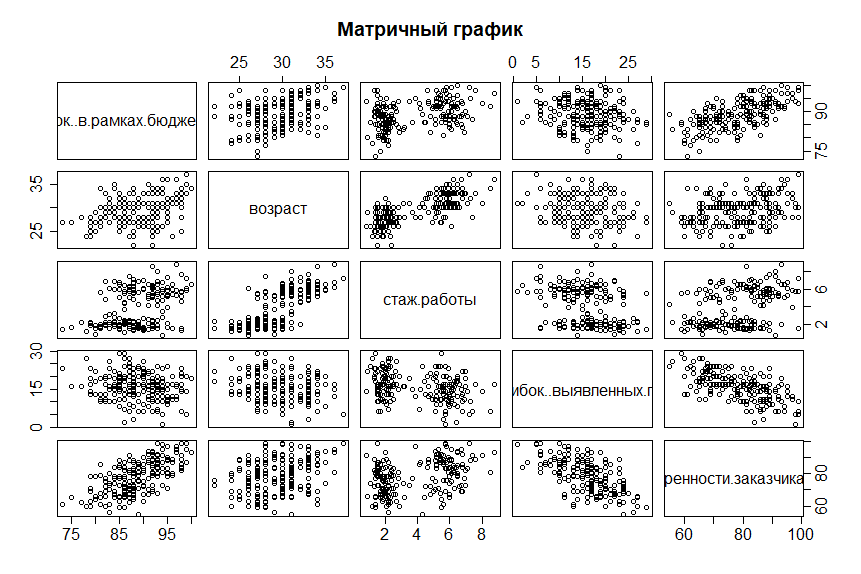


Рисунок – Матричный график для всех количественных признаков

Из матричного графика видно, что во всех зависимостях, где присутствует признак «качество документирования» данные никак не коррелируют между собой. Зависимости, связанные со стажем работы, позволяют разделить пользователей на две группы. Наблюдается зависимость между % выполнения работ в срок, степенью удовлетворенности заказчика и количеством найденных ошибок.

## Корреляционный анализ данных

Проверим гипотезу о наличии связи между полом и степенью удовлетворенности заказчика (качественная оценка) в каждой группе.

Примем уровень значимости . Нулевая гипотеза: пол разработчика и степень удовлетворенности заказчика не имеют связи.

В результате использования критерия χ2 (Chi-квадрат) и Фишера получены следующие данные (табл. 4 и 5).

Таблица 4. Результаты расчетов на основе критерия Chi-квадрат.

|  |  |  |  |
| --- | --- | --- | --- |
|  | X-квадрат | Df | p-значение |
| Группа 1 | 4.5033 | 2 | 0.1052 |
| Группа 2 | 1.468 | 2 | 0.48 |

В обеих группах p > 0.05, нулевая гипотеза об отсутствии связи не отвергается, т.е. нет связи между полом разработчика и степенью удовлетворенности заказчика.

Таблица 5. – Результаты расчетов на основе критерия Фишера.

|  |  |  |
| --- | --- | --- |
|  | p-значение | альтернативная гипотеза |
| Группа 1 | 0.1101 | двусторонняя |
| Группа 2 | 0.5318 | двусторонняя |

В обеих группах p > 0.05, нулевая гипотеза об отсутствии связи не отвергается.

Оценим связь между качественной переменной «степень удовлетворенности заказчика (качественная оценка)» и количественной переменной «процент разработок, выполненных в срок в рамках бюджета» на основе использования однофакторного дисперсионного анализа (ANOVA).

Примем уровень значимости . Нулевая гипотеза: процент разработок, выполненных в срок в рамках бюджета и степень удовлетворенности заказчика не имеют связи.

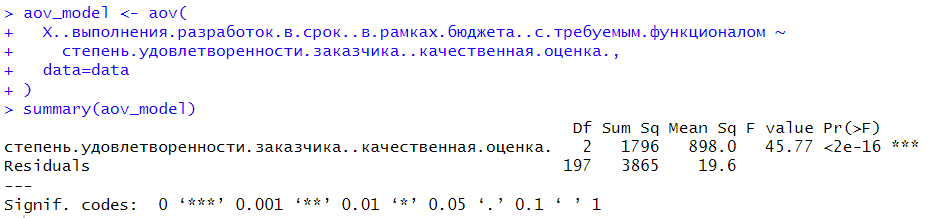


Рисунок – Тест ANOVA

Вычисленное значение p значительно меньше уровня значимости 0.05, нулевая гипотеза отвергается. Между процентом разработок, выполненных в срок в рамках бюджета и степенью удовлетворенности заказчика есть прямая связь.

Рассчитаем коэффициенты корреляции Пирсона, Спирмена, Кендалла для количественных признаков в каждой группе (рис. 12-17).

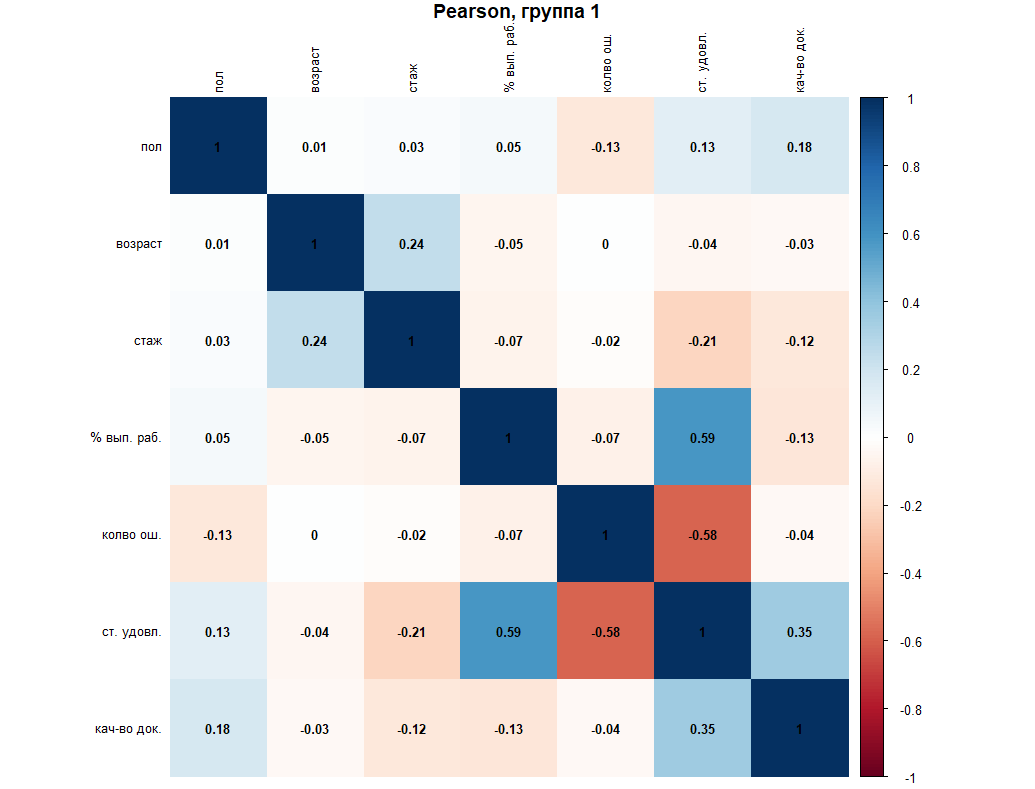


Рисунок – Pearson, группа 1

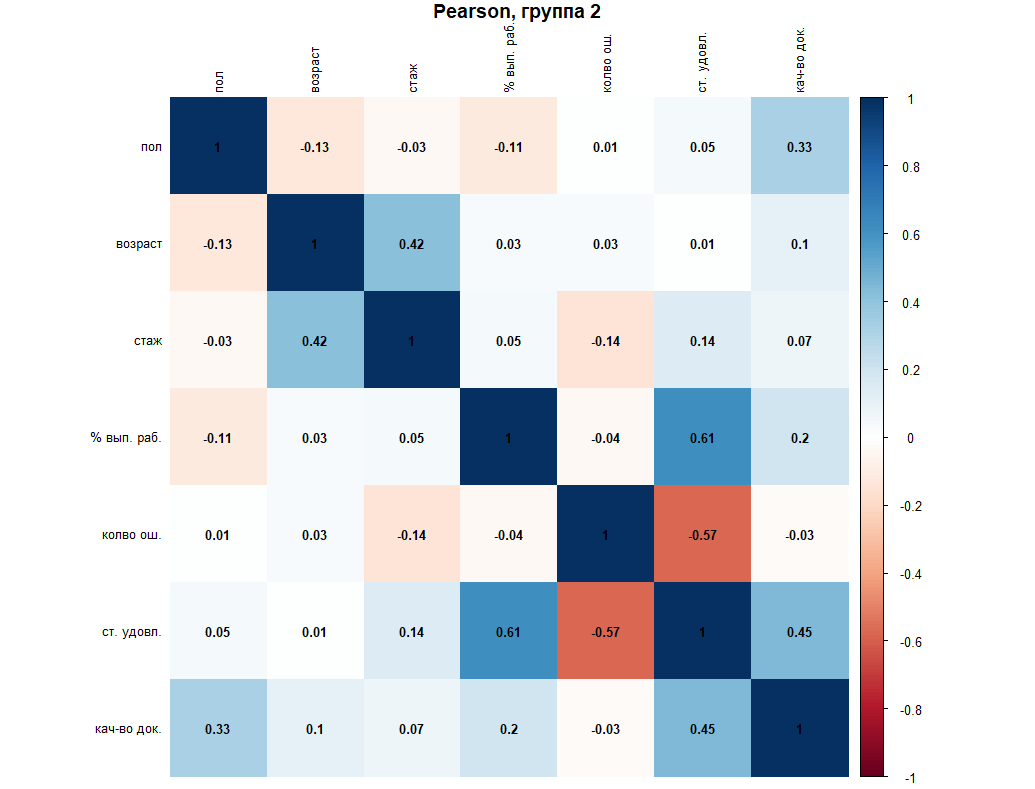


Рисунок – Pearson, группа 2

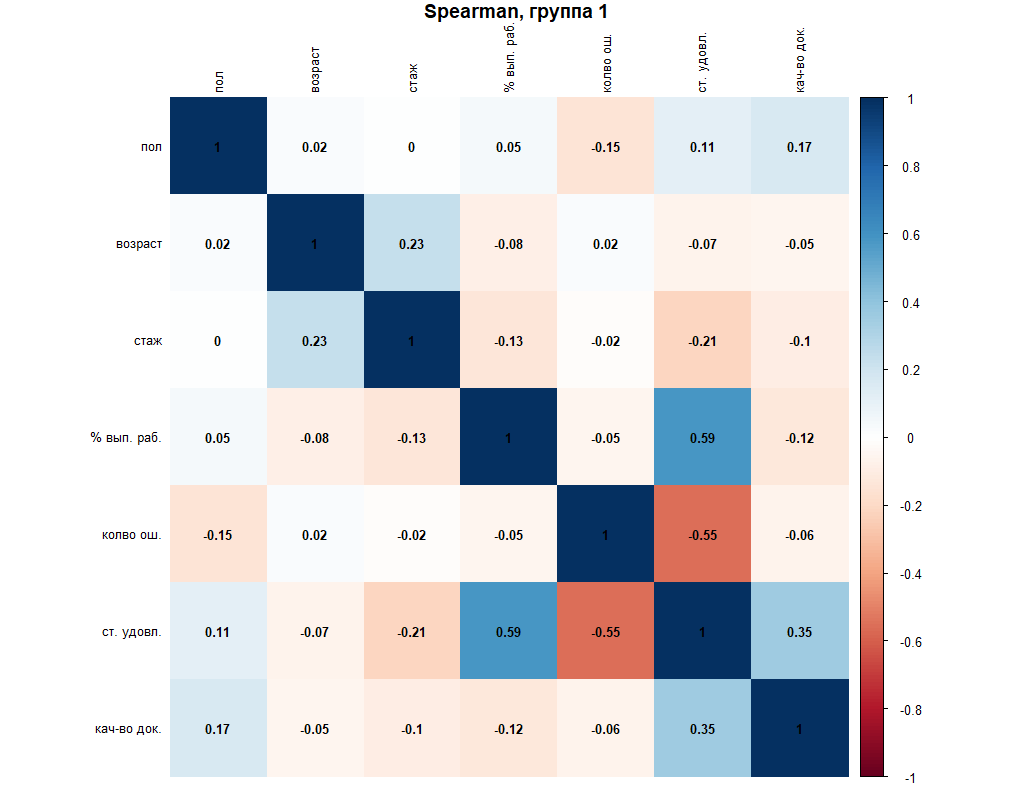


Рисунок – Spearman, группа 1

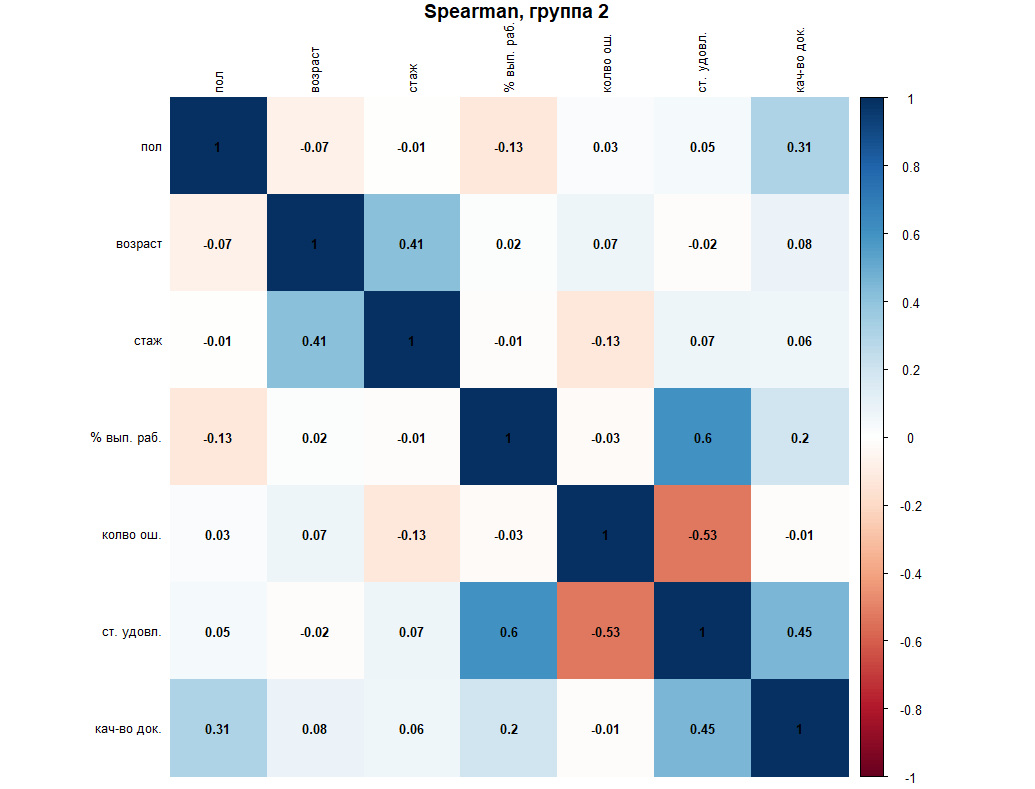


Рисунок – Spearman, группа 2

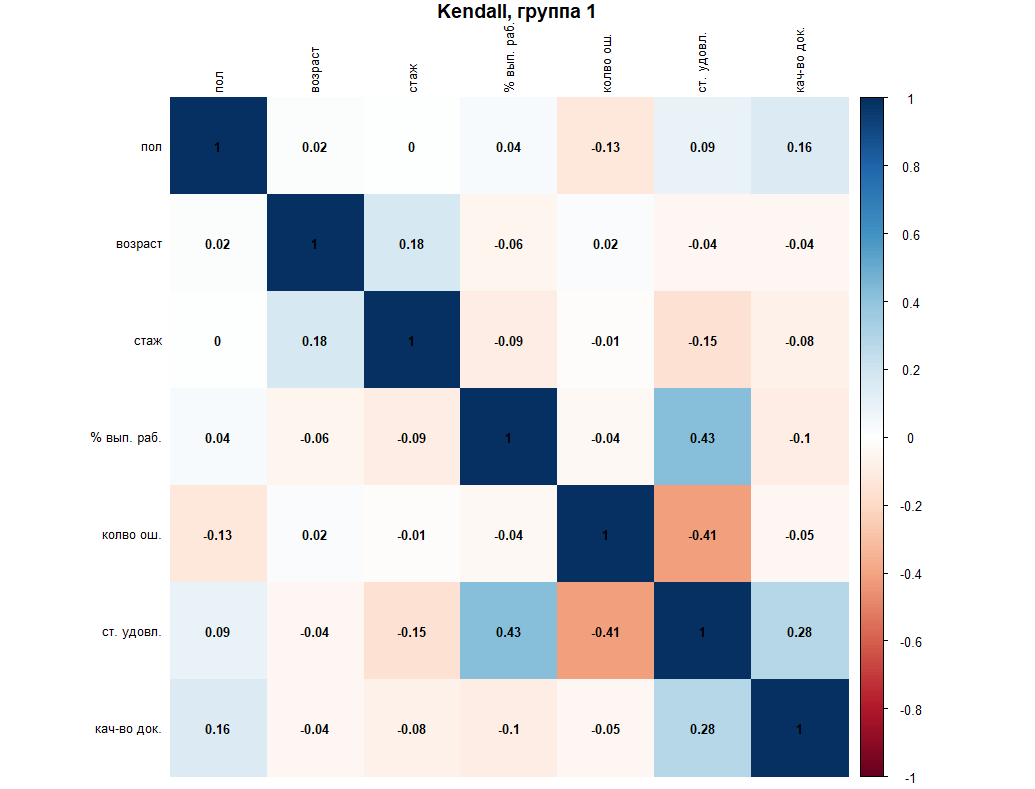


Рисунок – Kendall, группа 1

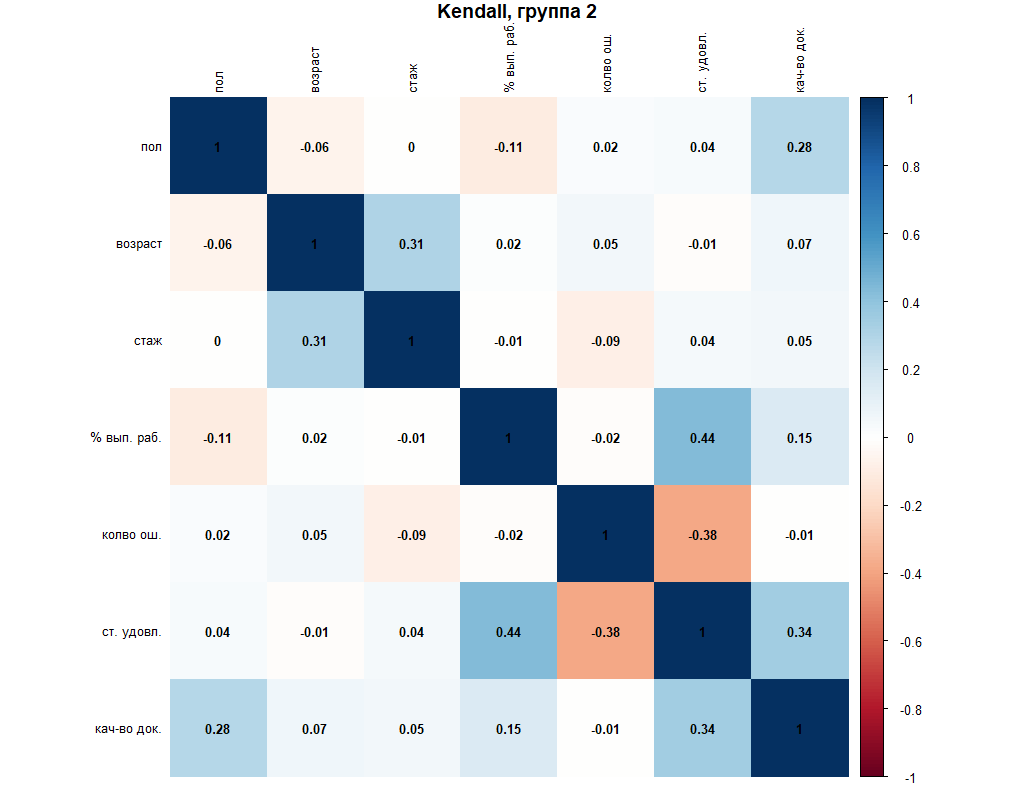


Рисунок – Kendall, группа 2

Чем ближе абсолютное значение коэффициента к 1, тем сильнее связь. Знак коэффициента говорит о направлении связи, положительный коэффициент – прямая связь, отрицательный – обратная связь.

Матрицы корреляции, полученные тремя способами, показывают одинаковое направление связи. Матрицы корреляции для разных групп мало отличаются между собой. Пол разработчика слабо влияет на другие переменные. Наблюдается незначительная прямая связь между полом и качеством документирования, эта связь усиливается во второй группе. Также присутствует средняя прямая связь между возрастом и стажем работы, процентом выполненных работ в срок и степенью удовлетворенности заказчика. Количество обнаруженных пользователем ошибок отрицательно влияет на степень удовлетворенности заказчика.

Оценим связь между количественными переменными «степень удовлетворенности заказчика (балльная оценка)» и «процент разработок, выполненных в срок в рамках бюджета» на основе расчета частного коэффициента корреляции для первой и второй групп.

Таблица 6 – Результаты расчетов частных коэффициентов корреляции.

|  |  |
| --- | --- |
|  | Частный коэффициент корреляции |
| Группа 1 | 0.8127962 |
| Группа 2 | 0.7379407 |

Между двумя переменными наблюдается сильная линейная прямая связь, во второй группе «сила» связи ниже.

## Вывод

В ходе выполнения работы были получены базовые навыки работы с языком программирования R и средой разработки R-studio. На примере решения конкретной задачи ИАД были исследованы зависимости между признаками программистов их результатами работы в крупной организации. Изучены средства R для проведения первичного разведочного анализа данных.

## Приложение

Исходный код:

library(psych) # describe

library(ggm)

library(corrplot)

# 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 1

# 1 группа – стаж менее 3,5

# 2 группа - стаж более 3,5

data <- read.csv("./var1lab1.csv", header=TRUE, sep=";")

data <- subset(data, select = -X.п.п) # drop column X.п.п

data$группа = as.factor(data$группа)

data$пол = as.factor(data$пол)

data$качество.документирования = as.factor(data$качество.документирования)

data$степень.удовлетворенности.заказчика..качественная.оценка. = as.factor(data$степень.удовлетворенности.заказчика..качественная.оценка.)

group1 <- subset(data, группа == 1)

group2 <- subset(data, группа == 2)

gender1 <- subset(data, пол == 1)

gender2 <- subset(data, пол == 2)

View(data)

#View(data[c(1,3,5)])

#View(data$пол[1:90])

#str(data)

find\_mode <- function(x) {

# function to compute unique values

# in vector

unq\_data <- unique(x)

# map values to its number of occurrences

map\_data <- match(x, unq\_data)

# table of the data with its values

tabulate\_data <- tabulate(map\_data)

# compute maximum value from data

max\_val <- max(tabulate\_data)

# plot it from table

unq\_data[tabulate\_data == max\_val]

}

find\_mode\_df <- function(x) {

print ("Мода для каждой колонки \n")

for (i in 2 : ncol(x)){

# calculating mode of ith column

mod\_val <- find\_mode(x[,i])

name <- colnames(x)[i]

print (c(name,": ", mod\_val), collapse=" ")

}

}

# 3. 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.

# mean - среднее значение

# min - минимальное значение

# max - максимальное значение

# sd - standard deviation - стандартное отклонение

# median - медиана, 50-й квантиль, 50% выборки имеют значение меньше, 50% больше

# trimmed - среднее значение после удаления некоторого %

# самых малых и больших значений из выборки

# mad - median absolute deviation, абс. отклонение медианы

# range - размах

# skew - ассиметрия распределения вероятности

# kurtosis - эксцесс, мера интенсивности выбросов относительно нормального распределения

# se - standard error, стандартная ошибка, мера неопределенности данного критерия

# первый квартиль - 25-й квантиль

# третий квартиль - 75-й квантиль

# мода - наиболее часто встречающееся значение

View(describe(data))

View(describe(group1))

View(describe(group2))

# квартили

View(summary(data))

View(summary(group1))

View(summary(group2))

print("Для всей выборки")

find\_mode\_df(data)

print("Для группы 1")

find\_mode\_df(group1)

print("Для группы 2")

find\_mode\_df(group2)

#plot(data.frame(data$количество.ошибок..выявленных.пользователем,data$стаж))

#boxplot(data$стаж)

# 4.

# 4.1 График рассеяния

plot(

data.frame(data$возраст,data$стаж.работы),

main="График рассеяния",

xlab = "Возраст", ylab = "Стаж работы",

pch = 10,

xaxp = round(c(min(data$возраст), max(data$возраст),20)),

yaxp = round(c(min(data$стаж.работы), max(data$стаж.работы),8))

)

# 4.2 Радиальный график для категориальных данных

drawPieCat <- function(column, main\_label) {

freqs <- rle(sort(column))

cat\_var <- factor(c(

rep(freqs$values, freqs$lengths)

))

cat <- table(cat\_var)

colname <- colnames(column)[0]

pie(

cat, radius=1,

main=main\_label,

col = hcl.colors(length(cat), "BluYl"),

clockwise=TRUE,

labels = paste0(round(100 \* cat/sum(cat), 2), "%")

)

legend(

"topright",

c(freqs$values),

cex=0.8, fill=hcl.colors(length(cat), "BluYl")

)

}

par(mfrow = c(1, 1))

drawPieCat(

data$степень.удовлетворенности.заказчика..качественная.оценка.,

"Степень удовлетворенности заказчика"

);

# 4.3

drawPieCat(

group1$степень.удовлетворенности.заказчика..качественная.оценка.,

"Степень удовлетворенности заказчика, группа 1"

);

drawPieCat(

group2$степень.удовлетворенности.заказчика..качественная.оценка.,

"Степень удовлетворенности заказчика, группа 2"

);

drawPieCat(

gender1$степень.удовлетворенности.заказчика..качественная.оценка.,

"Степень удовлетворенности заказчика, пол 1"

);

drawPieCat(

gender2$степень.удовлетворенности.заказчика..качественная.оценка.,

"Степень удовлетворенности заказчика, пол 2"

);

# 4.4 Стобиковая диаграмма

par(mfrow = c(1, 1))

qualityInGroup1 = mean(group1$степень.удовлетворенности.заказчика..балльная.оценка.)

qualityInGroup2 = mean(group2$степень.удовлетворенности.заказчика..балльная.оценка.)

qualityInGender1 = mean(gender1$степень.удовлетворенности.заказчика..балльная.оценка.)

qualityInGender2 = mean(gender2$степень.удовлетворенности.заказчика..балльная.оценка.)

quality = c(qualityInGroup1, qualityInGroup2, qualityInGender1, qualityInGender2)

barplot(

quality,

main="Средняя степень удовлетворенности заказчика (балльная оценка)",

names=c("Группа 1", "Группа 2", "Пол 1", "Пол 2"),

col=hcl.colors(length(cat), "BluYl"),

xlab="", ylab="Cтепень удовлетворенности заказчика (балльная оценка)"

)

# 4.5

par(mfrow = c(1, 1))

boxplot(

X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом ~ группа,

main="% выполнения разработок в срок", ylab="% выполнения разработок в срок",

data=data,

ylim = c(

min(

data$X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом)-1,

max(

data$X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом)+1

)

)

# 4.6

par(mfrow = c(3, 2))

hist(

data$X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом,

main="% выполнения разработок в срок c требуемым функционалом",

xlab=""

)

hist(

data$возраст,

main="Возраст",

xlab=""

)

hist(

data$стаж.работы,

main="Стаж работы",

xlab=""

)

hist(

data$количество.ошибок..выявленных.пользователем,

main="Количество ошибок, выявленных пользователем",

xlab=""

)

hist(

data$степень.удовлетворенности.заказчика..балльная.оценка.,

main="Степень удовлетворенности заказчика (балльная оценка)",

xlab=""

)

# 4.7

par(mfrow = c(1, 1))

pairs(

~X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом +

возраст +

стаж.работы +

количество.ошибок..выявленных.пользователем +

степень.удовлетворенности.заказчика..балльная.оценка.,

data=data, main="Матричный график",

cex.labels=1.5,

cex.axis=1.5

)

# 5.1

# Chi squared

print("Группа 1")

print(chisq.test(

table(

group1$пол,

group1$степень.удовлетворенности.заказчика..качественная.оценка.

),

correct=TRUE

))

print("Группа 2")

print(chisq.test(

table(

group2$пол,

group2$степень.удовлетворенности.заказчика..качественная.оценка.

),

correct=TRUE

))

# Fisher

print("Группа 1")

print(fisher.test(

table(

group1$пол,

group1$степень.удовлетворенности.заказчика..качественная.оценка.

)

))

print("Группа 2")

print(fisher.test(

table(

group2$пол,

group2$степень.удовлетворенности.заказчика..качественная.оценка.

)

))

# 5.2 ANOVA

# различаются ли средние значения между группами

aov\_model <- aov(

X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом ~

степень.удовлетворенности.заказчика..качественная.оценка.,

data=data

)

summary(aov\_model)

# https://www.scribbr.com/statistics/anova-in-r/

#df = the number of levels in the variable minus 1

# residuals df = the total number of observations minus one and minus

# the number of levels in the independent variables

# Pr(>F) - p value. This shows how likely it is that the F value calculated

# from the test would have occurred if the null hypothesis of no difference

# among group means were true.

# у нас p = <2e-16 < 0.05, что говорит о наличии связи между группой и

# cтепенью удовлетворенности заказчика

# 5.3 Рассчет коэффициентов корреляции, 5.5

# Линейный коэффициент корреляции Пирсона (Pearson product moment correlation)

# отражает степень линейной связи между двумя количе-ственными переменными.

# Коэффициент ранговой корреляции Спирмана (Spearman’s Rank Order correlation)

# – мера связи между двумя ранжиро-ванными переменными.

# Тау Кендалла – также непараметрический показатель ранговой корреляции.

drawMatrix <- function(N, main\_label) {

par(mfrow = c(1, 1))

names <- c(

'пол',

'возраст',

'стаж',

'% вып. раб.',

'колво ош.',

'ст. удовл.',

'кач-во док.'

)

rownames(N) <- names

colnames(N) <- names

corrplot(

N, method="color",

addCoef.col = 1,

tl.cex = 0.8, number.cex = 0.8,

tl.col="black",

main=main\_label,

mar=c(0,0,1,0)

)

}

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

M1 <- subset(M, группа==1, select= -группа)

M2 <- subset(M, группа==2, select= -группа)

#View(M1)

print("Группа 1")

N1 <- cor(M1,use="pairwise.complete.obs", method="pearson")

N2 <- cor(M1,use="pairwise.complete.obs", method="spearman")

N3 <- cor(M1,use="pairwise.complete.obs", method="kendall")

#print("Pearson")

#print(N1)

#print("Spearman")

#print(N2)

#print("Kendall")

#print(N3)

drawMatrix(N1, "Pearson, группа 1")

drawMatrix(N2, "Spearman, группа 1")

drawMatrix(N3, "Kendall, группа 1")

print("Группа 2")

N1 <- cor(M2,use="pairwise.complete.obs", method="pearson")

N2 <- cor(M2,use="pairwise.complete.obs", method="spearman")

N3 <- cor(M2,use="pairwise.complete.obs", method="kendall")

#print("Pearson")

#print(N1)

#print("Spearman")

#print(N2)

#print("Kendall")

#print(N3)

drawMatrix(N1, "Pearson, группа 2")

drawMatrix(N2, "Spearman, группа 2")

drawMatrix(N3, "Kendall, группа 2")

# 5.4

#View(M1)

#View(M2)

#M$степень.удовлетворенности.заказчика..балльная.оценка.

#M$X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом

print(pcor(c(4,6, 1,2,3,5,7), cov(M1)))

print(pcor(c(4,6, 1,2,3,5,7), cov(M2)))

#

print("Pearson test")

print(cor.test(

M1$степень.удовлетворенности.заказчика..балльная.оценка.,

M1$X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом,

method="pearson"

))

print(cor.test(

M2$степень.удовлетворенности.заказчика..балльная.оценка.,

M2$X..выполнения.разработок.в.срок..в.рамках.бюджета..с.требуемым.функционалом,

method="pearson"

))