Министерство науки и высшего образования Российской Федерации

Новосибирский государственный технический университет

Кафедра Вычислительной техники



Лабораторная работа №1

По дисциплине «Методы анализа данных»

По теме

«Методы первичного разведочного анализа в R».

Факультет: АВТ

Группа: АВТ-819

Студент: Ванин К.Е.

Преподаватель: Пешков А.В.

Новосибирск

2022

**Цель работы**

- получить базовые навыки работы в среде R;

- изучить средства R для проведения первичного разведочного анали-за данных (методы визуализации, описательной статистики, корре-ляционного анализа данных) на примере решения конкретной зада-чи 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, первая и третья 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.

5. Проведите корреляционный анализ данных.

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

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

5.3. 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.

5.4. Графически представьте матрицы коэффициентов корреляции. Оцените статистическую значимость связи. Сделайте выводы о силе и направлении связи в терминах решаемой задачи.

6. Сделайте выводы по работе в произвольной форме

**Постановка задачи:**

Исследуются покупатели Интернет-магазина. Было опрошено 200 клиентов, каждому анкетируемому предлагалось ответить на следующие вопросы:

* возраст;
* пол (1-м, 2-ж);
* количество покупок за год;
* средняя стоимость покупок за год;
* среднее число страниц, просмотренных за визит;
* количество обращений в службу поддержки за год;
* степень удовлетворенности услугами;
* степень активности (участие в Интернет-опросах)  (1 – никогда, 2 – редко, 3 – часто, 4 – постоянно)

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

Анкетируемые разбиты на две группы в зависимости от возраста: 1 группа – возраст менее 30 лет, 2 группа – возраст более 30 лет

**Ход работы:**

До начала работы с массивом данных необходимо загрузить таблицу в среду обработки:

data <- read.table("C:\\Users\\рбт\\Desktop\\lr1\_v8.csv", header=TRUE, sep=";")

View(data)

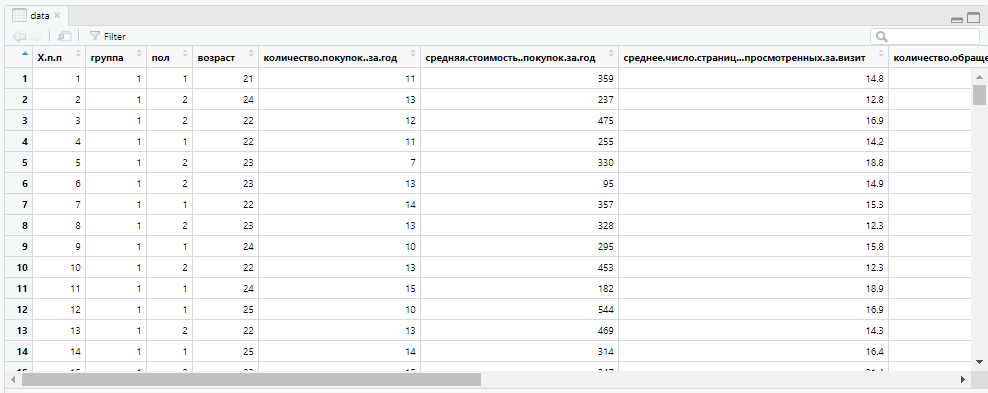


Рис.1 – фрагмент таблицы БД

**1. Расчёт основных статистических характеристик для всей выборки.**

Рассчитайте основные статистические характеристики по количественным данным, отдельно для первой и второй групп и для всей выборки:

* минимальное и максимальное значения выборки
* выборочное среднее
* выборочное среднеквадратичное (стандартное) отклонение
* медиана и мода выборки
* 1-ый и 3-ий квартили выборки
* коэффициенты асимметрии и эксцесс

#Статистика

summary(data)

#Функция

mystats(data$возраст)

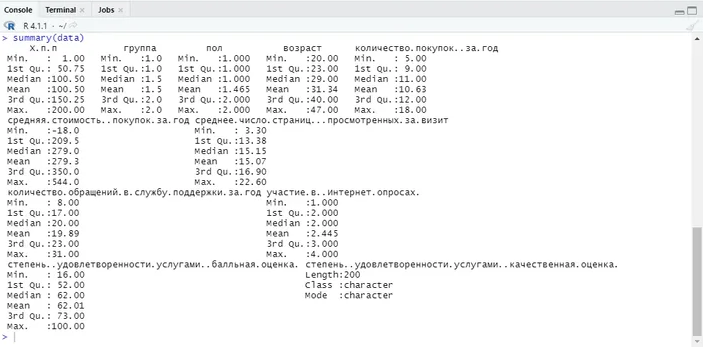


Рис.2 – Результат ввода функции **summary**

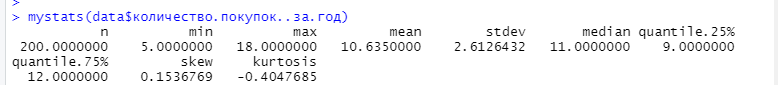


Рис. 3 – Результат ввода функции **mystat**

Таблица Результаты расчетов статистических характеристик по количественным данным

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | среднее | min | max | стандартное отклонение | квартили | | медиана | мода | асимметрия | эксцесс |
| Q1 | Q3 |
| **Кол-во покупок за год** | 10.6 | 5 | 18 | 2.61 | 9 | 12 | 11 | 12 | 0.1536769 | -0.404768 |
| **Средняя стоимость покупок за год** | 279.285 | -18 | 544 | 104.84 | 209.5 | 350 | 279 | 318 | -0.09503962 | -0.177028 |
| **Кол-во обращений в службу поддержки за год** | 19,895 | 8 | 31 | 4,26 | 17 | 23 | 20 | 23 | -0.0704034 | -0.359436 |

Таблица Результаты расчетов статистических характеристик по количественным данным для 1 группы

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Группа 1 | среднее | min | max | стандартное отклонение | квартили | | медиана | мода | асимметрия | эксцесс |
| Q1 | Q3 |
| **Кол-во покупок за год** | 12,2 | 7 | 18 | 2,23 | 11 | 14 | 12 | 12 | 0.01507203 | -0.255385 |
| **Средняя стоимость покупок за год** | 323,58 | 95 | 544 | 88,91 | 261 | 375 | 332 | 375 | -0.09406339 | -0.099927 |
| **Кол-во обращений в службу поддержки за год** | 19,9 | 11 | 30 | 4,27 | 16,75 | 23 | 20 | 21 | 0.03828048 | -0.829112 |

Таблица Результаты расчетов статистических характеристик по количественным данным для 2 группы

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Группа 2 | среднее | min | max | стандартное отклонение | квартили | | медиана | мода | асимметрия | эксцесс |
| Q1 | Q3 |
| **Кол-во покупок за год** | 9 | 5 | 13 | 1,9 | 8 | 10 | 9 | 10 | -0.04034057 | -0.836353 |
| **Средняя стоимость покупок за год** | 234.99 | -18 | 504 | 101.1 | 172 | 300 | 230,5 | 203 | 0.14709850 | 0.065872 |
| **Кол-во обращений в службу поддержки за год** | 19.89 | 8 | 31 | 4.26 | 17 | 23 | 20 | 20 | -0.17905960 | 0.061755 |

Полученные результаты позволяют сделать следующие выводы:

Средние значения по количеству покупок и стоимости за одну покупку за год больше у первой группы по сравнению со второй. Значение медианы также значительно различается между группами, у первой группы этот показатель выше по количеству покупок и стоимости за одну покупку за год. Значения по показателю количество обращений в службу поддержки практически одинаково как во всей выборке, так и в равнении между группами. Можно сказать, что в первой группе количество обращений в службу поддержки будет чуть больше.

Стоит отметить, что во второй группе средняя стоимость покупок в год имеет минимальное отрицательное значение, из-за этого значения во второй группе будут значительно ниже по покупкам, чем во второй группе.

Доминирующий отрицательный коэффициент эксцесса у большинства переменных говорит о том, что распределение является более «пологим» относительно нормального распределения.

Положительный коэффициент асимметрии у переменной «общее кол-во покупок за год» означает, что в распределении чаще встречаются значения меньше среднего, а отрицательный коэффициент у переменных «ежемесячный платеж»; «Кол-во обращений в службу поддержки за год» - чаще встречаются значения больше среднего.

**2. Графический анализ данных**

**диаграмма рассеяния по двум количественным признакам;**

plot(data.frame(data$возраст,data$количество.покупок..за.год), type="b", col="purple", lty=0, pch=19, lwd=1, main="Диаграмма рассеяния", sub="Васюков, Жирнов", xlab="Возраст", ylab="Кол-во покупок за год", xlim=c(19, 47), ylim=c(5, 20))

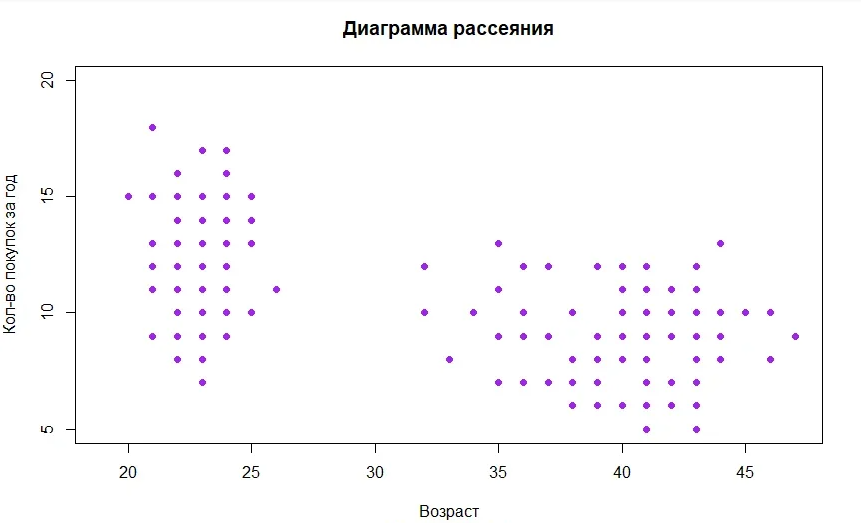


Рис. 4 – Диаграмма рассеяния индекса покупок за год, в зависимости от возраста

Из данной диаграммы рассеяния можно увидеть, что покупатели в возрасте от 20 до 25 лет совершают больше покупок в год (в среднем около 15), чем люди в возрасте от 35 до 45 лет (в среднем около 10 покупок). Для анализа количества покупок людей в возрасте от 25 до 35 недостаточно данных.

**- радиальная диаграмма по качественному признаку;**

x <- data$степень..удовлетворенности.услугами..качественная.оценка.

t <-data.frame(table(x))$Freq

piedata <- round(100\*t/sum(t), 1)

pie (t,piedata,radius=1, main="Радиальная диаграмма", labels=levels(x))

1 – высокая

2 – низкая

3 - средняя



Рис. 5 – радиальная диаграмма

Больше половины покупателей ставят низкую оценку качеству услуг.

**- категориальная радиальная диаграмму по одному из качественных признаков**

par(mfrow = c(1, 2))

x <- data$степень..удовлетворенности.услугами..качественная.оценка.[data$возраст<30]

t <-data.frame(table(x))$Freq

piedata <- round(100\*t/sum(t), 1)

pie(t,piedata,radius=1, main="Люди младше 30 лет", labels=levels(x))

x <- data$степень..удовлетворенности.услугами..качественная.оценка.[data$возраст>30]

t <-data.frame(table(x))$Freq

piedata <- round(100\*t/sum(t), 1)

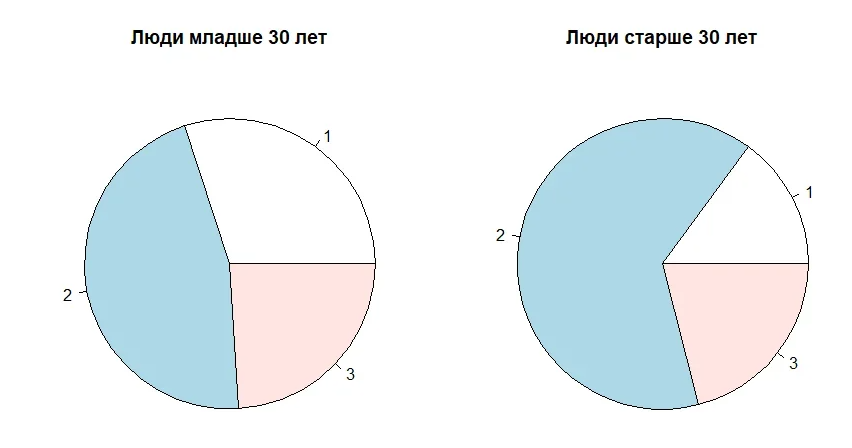
pie(t,piedata,radius=1, main="Люди старше 30 лет", labels=levels(x)) 

Рис. 6 – Круговые диаграммы по оценке качества услуги для разных групп

Из данных двух диаграмм видно, что люди младше 30 лет реже ставят низкую оценку (~45%), чем люди старше 30 лет (~65%). Высоких оценок больше у людей младше 30 лет, средних оценок ближе к одинаковому значению.

**- категориальную столбиковую диаграмму по одному из количественных признаков**

barplot(table(data$пол[data$количество.покупок..за.год]), col="purple", xlab="Пол", ylab="Количество покупок", main="Количество покупок за год в зависимости от пола")

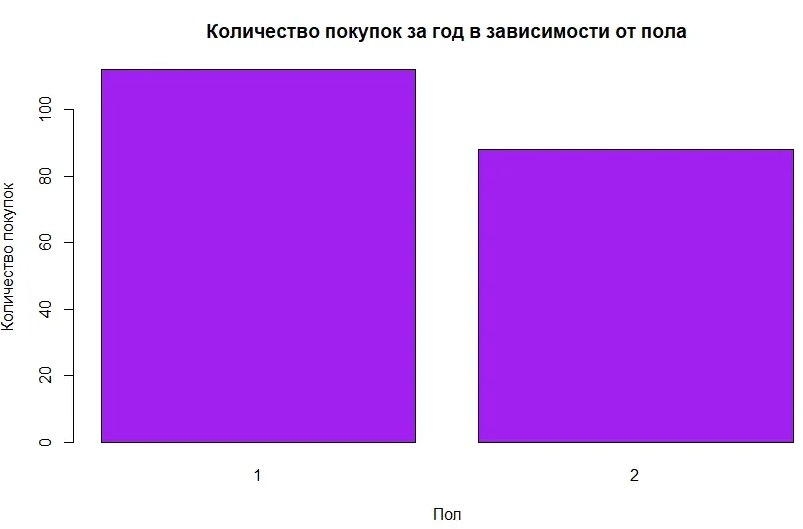


Рис. 7 – Столбчатая диаграмма

С помощью данной диаграммы можно заключить, что мужчины совершают больше покупок, нежели женщины. У мужчин ~110 покупок в год, у женщин ~90 покупок в год.

**- диаграмму размаха для одного из количественных признаков**

boxplot(data$количество.обращений.в.службу.поддержки.за.год[data$возраст<30],data$количество.обращений.в.службу.поддержки.за.год[data$возраст>30], main="Диаграмма количества обращений в службу поддержки за год",names=c("Люди младше 30","Люди старше 30"))

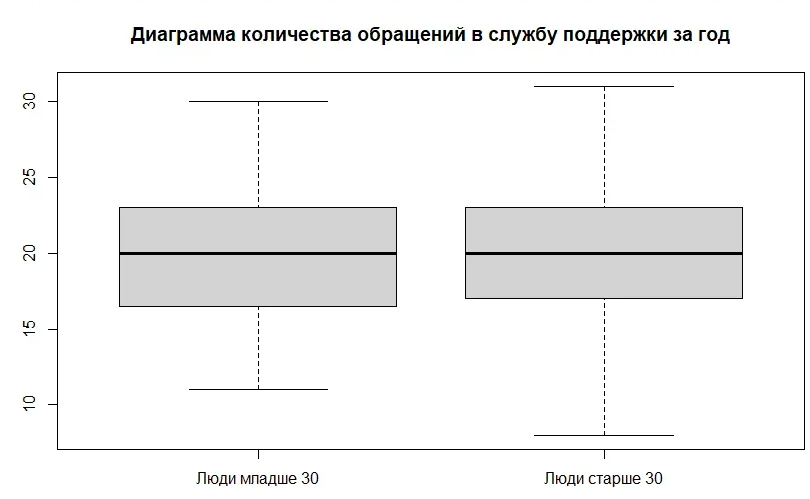


Рис. 8 – диаграмма размаха по количеству обращений в службу поддержки

По данной диаграмме можно видеть, что средний процент обращений в службу поддержки почти одинаковый. Люди старше 30 лет обращаются в службу поддержки чуть чаще людей младше 30 лет. Так же нижний кварталь у людей старше 30 лет ниже.

**- гистограммы для всех количественных признаков на одном графике;**

par(mfrow=c(2,4))

hist(data$количество.покупок..за.год, freq=FALSE, breaks=12, main="Количество покупок за год",xlab="Количество покупок")

hist(data$количество.обращений.в.службу.поддержки.за.год, freq=FALSE, breaks=12, main="Обращения за год",xlab="Количество")

hist(data$среднее.число.страниц...просмотренных.за.визит, freq=FALSE, breaks=12, main="Просмотренные страницы",xlab="за визит")

hist(data$средняя.стоимость..покупок.за.год, freq=FALSE, breaks=12, main="Средняя стоимость покупок",xlab="за год")

hist(data$возраст, freq=FALSE, breaks=12, main="Возраст",xlab="-")

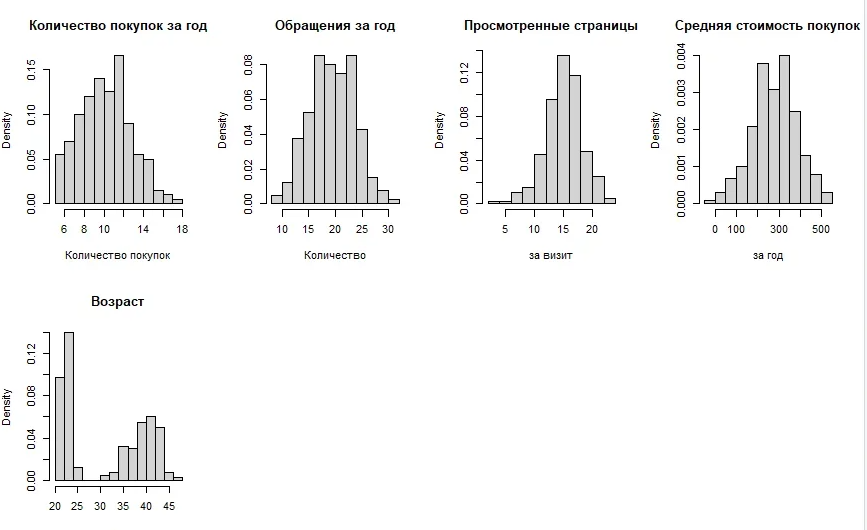


Рис. 9 – Гистограммы распределения количественных признаков общей выборки

**- матричный график по всем количественным переменным.**

pairs(~data$количество.покупок..за.год+data$количество.обращений.в.службу.поддержки.за.год+data$возраст)

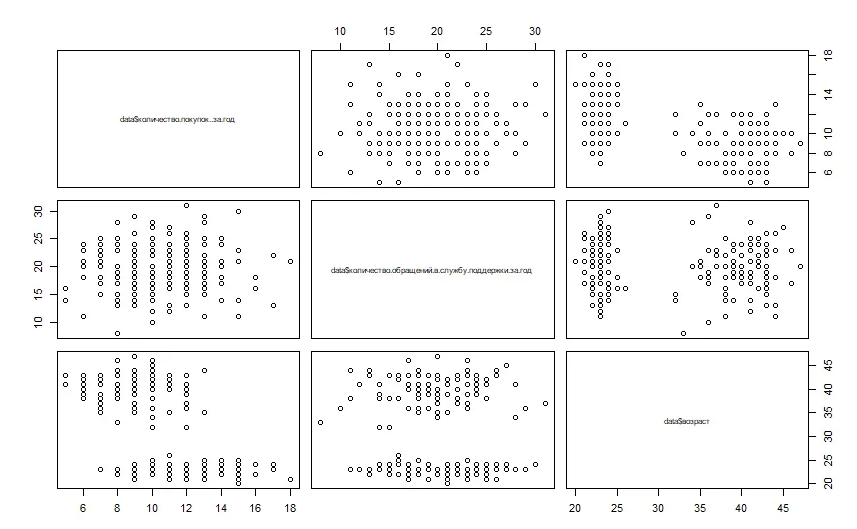


Рис. 9 – Матричная диаграмма рассеяния пар количественных признаков

**3. Корреляционный анализ данных**

**Оценка степени взаимосвязи между качественными переменными на основе критериев Chi-квадрат и Фишера.**

twoQualitativesGr1 = table(group1[, c("возраст", "степень..удовлетворенности.услугами..качественная.оценка.")])

twoQualitativesGr2 = table(group2[, c("возраст", "степень..удовлетворенности.услугами..качественная.оценка.")])

print("Хи-квадрат:")

chisq.test(twoQualitativesGr1, correct=TRUE)

chisq.test(twoQualitativesGr2, correct=TRUE)

print("Критерий Фишера:")

fisher.test(twoQualitativesGr1)

fisher.test(twoQualitativesGr2)

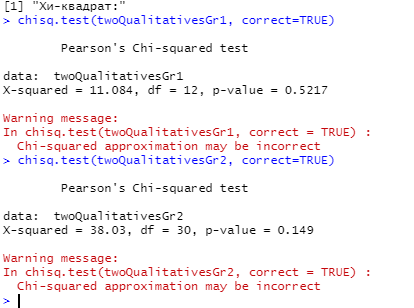


Рис. 10 – Матрица значений критерия хи-квадрат для пар качественных признаков

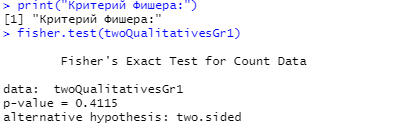
****

Рис. 11 – Степень взаимосвязи между качественными переменами

**Оценка степени взаимосвязи между количественными переменными на основе расчета коэффициентов корреляции Пирсона, Спирмена, Кендалла**

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

N1<-cor(M,use="pairwise.complete.obs")

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

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

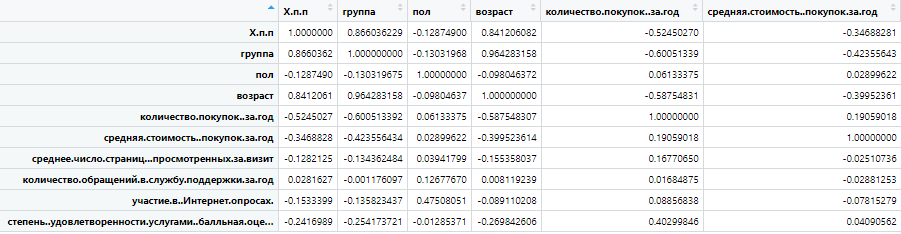


Рис. 12 – Матрица коэффициентов корреляции, полученных методом Пирсона

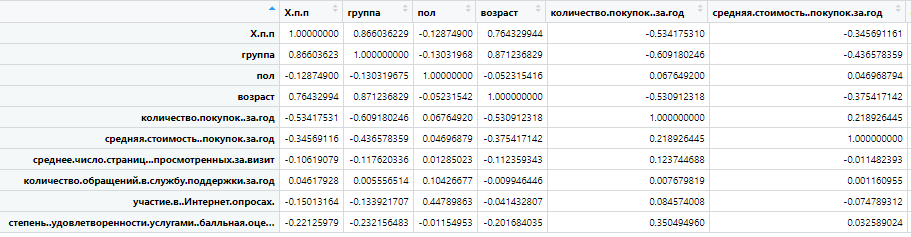


Рис. 13 – Матрица коэффициентов корреляции, полученных методом Спирмена



Рис. 14 – Матрица коэффициентов корреляции, полученных методом Кендалла

**Оценка степени взаимосвязи между двумя количественными переменными (для которых коэффициент корреляции Пирсона максимален по модулю) на основе расчета частного коэффициента корреляции**

install.packages("ggm")

library(ggm)

M1 <- data.unlist(lapply(data1, is.numeric))]

pcor(c(4,6,1,2,3,5,7,8,9), cov(M1))



Рис. 15 – Степень взаимосвязи между двумя количественными переменными

**Графическое представление матриц коэффициентов корреляции. Оценка статистической значимости связи.**

install.packages("corrplot")

library(corrplot)

corrplot(N11, 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

)

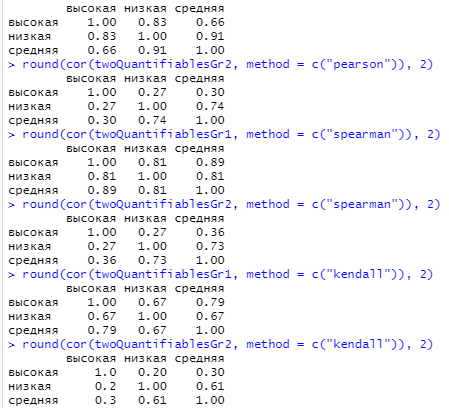


Рис. 16 – Оценка статистической значимости связи

**Вывод:**

В ходе выполнения лабораторной работы были получены базовые навыки работы в среде R, изучены средства R для проведения первичного разведочного анализа данных (методы визуализации, описательной статистики, корреляционного анализа данных) на примере решения конкретной задачи ИАД (интеллектуального анализа данных).

Так же из полученных диаграмм можем наблюдать, что: люди младше 30 лет чаще совершают покупки. В свою очередь, люди старше 30 лет чаще обращаются в службу поддержки. Из диаграммы покупок по полу видно, что мужчины совершают больше покупок. Средняя стоимость покупок равна ~300 денег.