ДИСЦИПЛИНА	Прикладные задачи математической статистики
ИНСТИТУТ	Институт искусственного интеллекта
КАФЕДРА	Технологий искусственного интеллекта
ГРУППЫ	
ВИДЫ УЧЕБНОГО МАТЕРИАЛА	Практические материалы для лабораторного занятия Тема - 4 Способы визуализации данных. Часть 2 (Лаб). Изучение пакетов R для визуализации взаимосвязей в данных
ПРЕПОДАВАТЕЛЬ	Бурцева Т.А.
СЕМЕСТР	1 семестр

Методические указания к выполнению заданий

Используя переменную t, выполните скрипт и задания в R.

Задание 1.

```
install.packages("car")
library(car)
pairs(t[2:3])
```

Ответ:

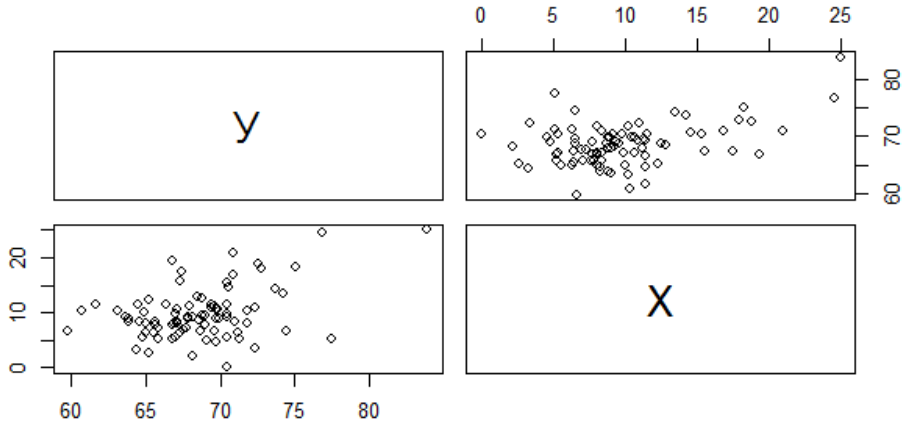


Рис.1

Задание 2

```
scatterplotMatrix( t[2:3])
```

Ответ:

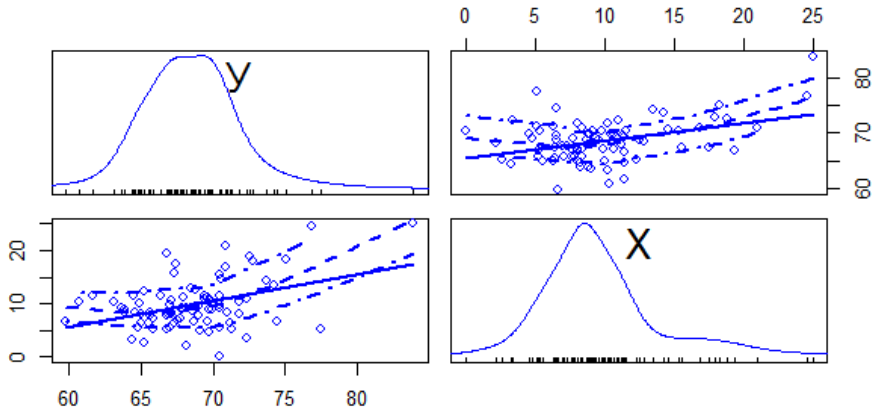


Рис. 2

Задание 3

```
scatterplotMatrix(t[2:3], regLine = FALSE, smooth = TRUE,
        main = "Correlations of parties' share",
        diagonal = list(method = "histogram"))
```

Ответ:

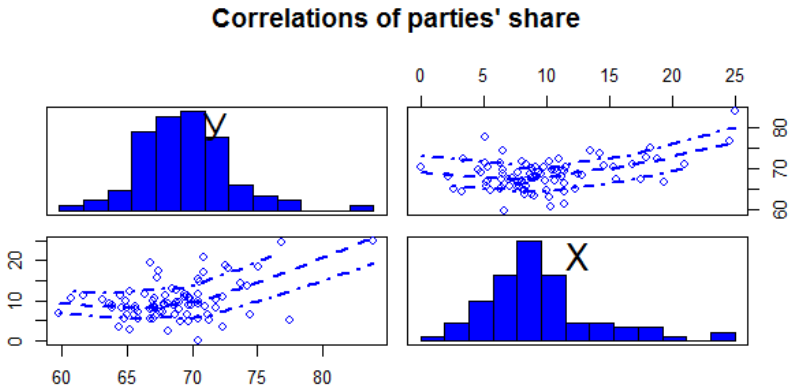


Рис.3

Задание 4

```
scatterplotMatrix(t[2:3], regLine = FALSE, smooth = TRUE,
        main = "Correlations of parties' share",
        diagonal = list(method = "boxplot"))
```

Ответ:

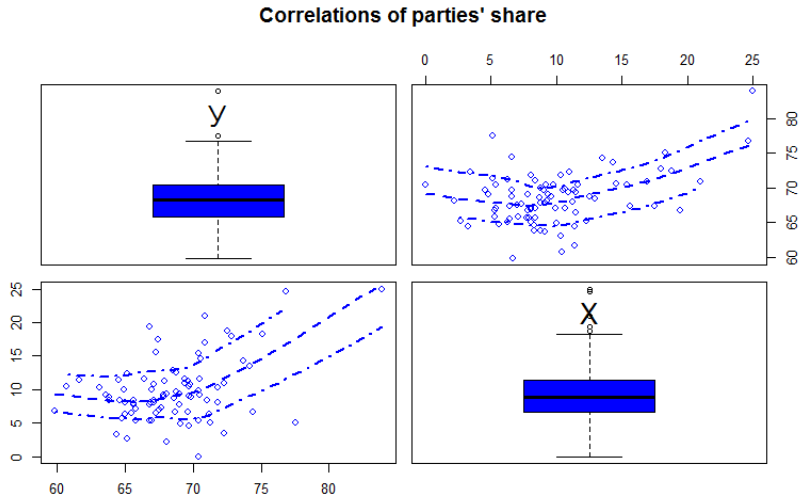
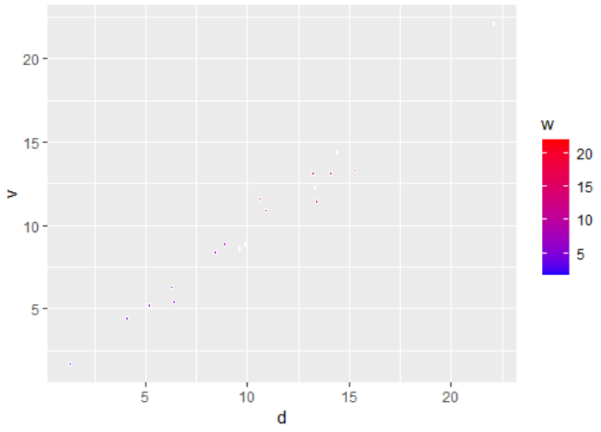


Рис. 4

Задание 5**.** Тепловая карта – это двумерное визуальное представление данных, которое использует цвета для отображения значения переменной. Тепловые карты часто используются для демонстрации корреляции различных факторов в наборе данных друг с другом. Пример создания тепловой карты:

```
d=c(6.4,13.3,4.1,1.3,14.1,10.6,9.9,9.6,15.3,22.1,13.4, 13.2,8.4,6.3,8.9,5.2,10.9,14.4)
v=c(5.4,12.3,4.4,1.7,13.1,11.6,8.9,8.6,13.3,22.1,11.4, 13.1,8.4,6.3,8.9,5.2,10.9,14.4)
data=as.data.frame(d)
data$v=v
data$w=v
library(ggplot2)
ggplot(data, aes(d, v)) +  geom_tile(aes(fill = w),colour = "white") +
  scale_fill_gradient(low = "blue",high = "red")
```

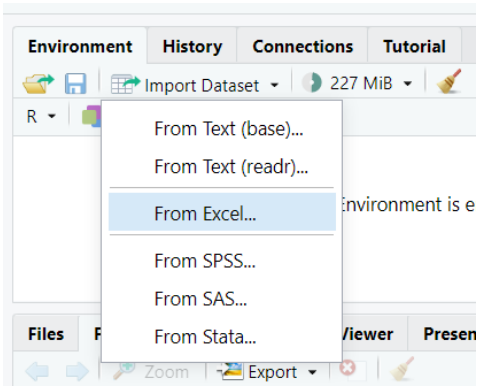
Ответ:

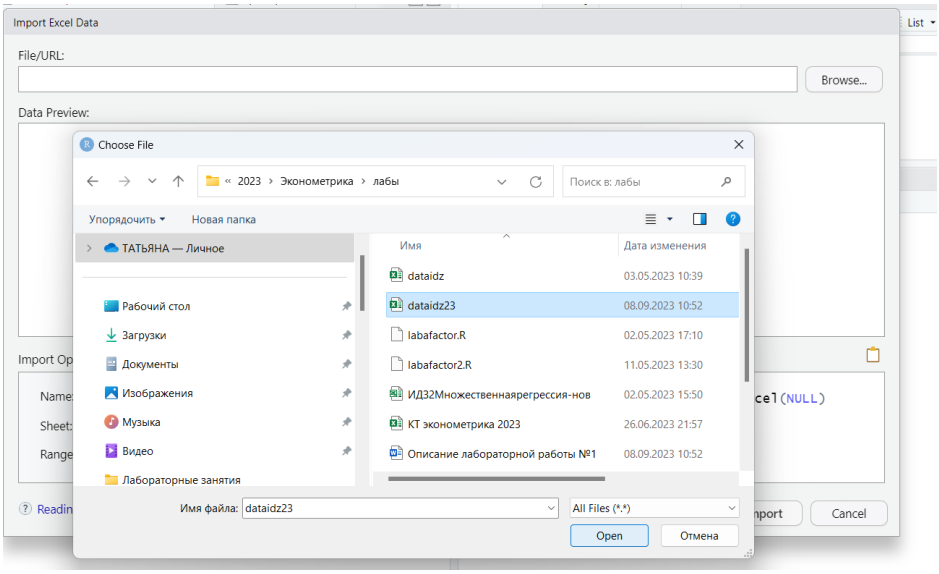


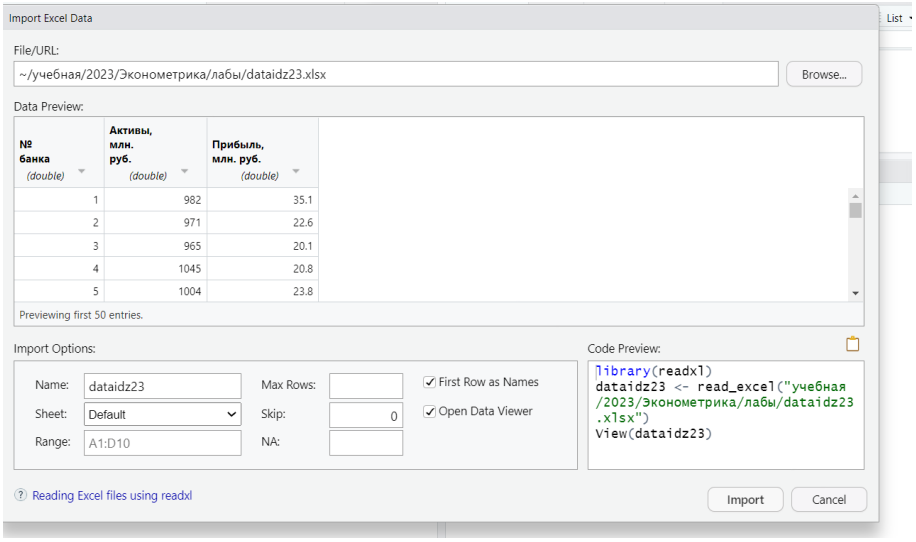
Задание 6. Построение моделей парной регрессии без и с учетом фактора с помощью языка статистической обработки данных R

Шаг. 1.

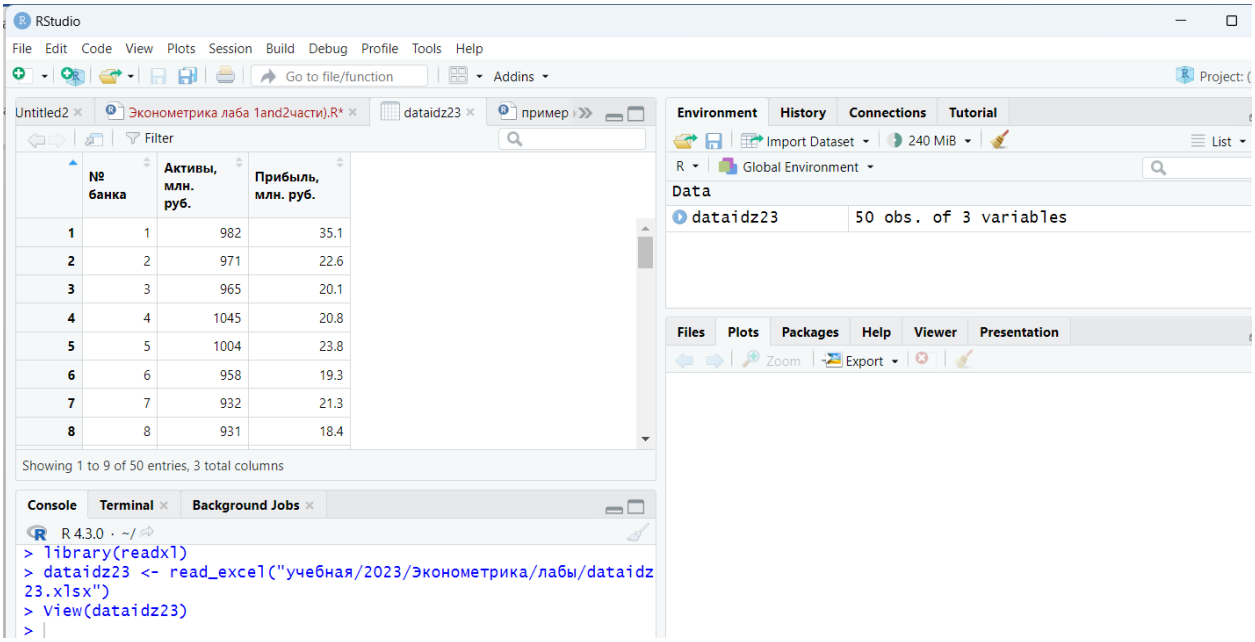
Загружаем таблицу данных для всех вариантов как объект данных в среду R-Studio, для чего используем кнопку Import Dataset, находим файл dataidz23:



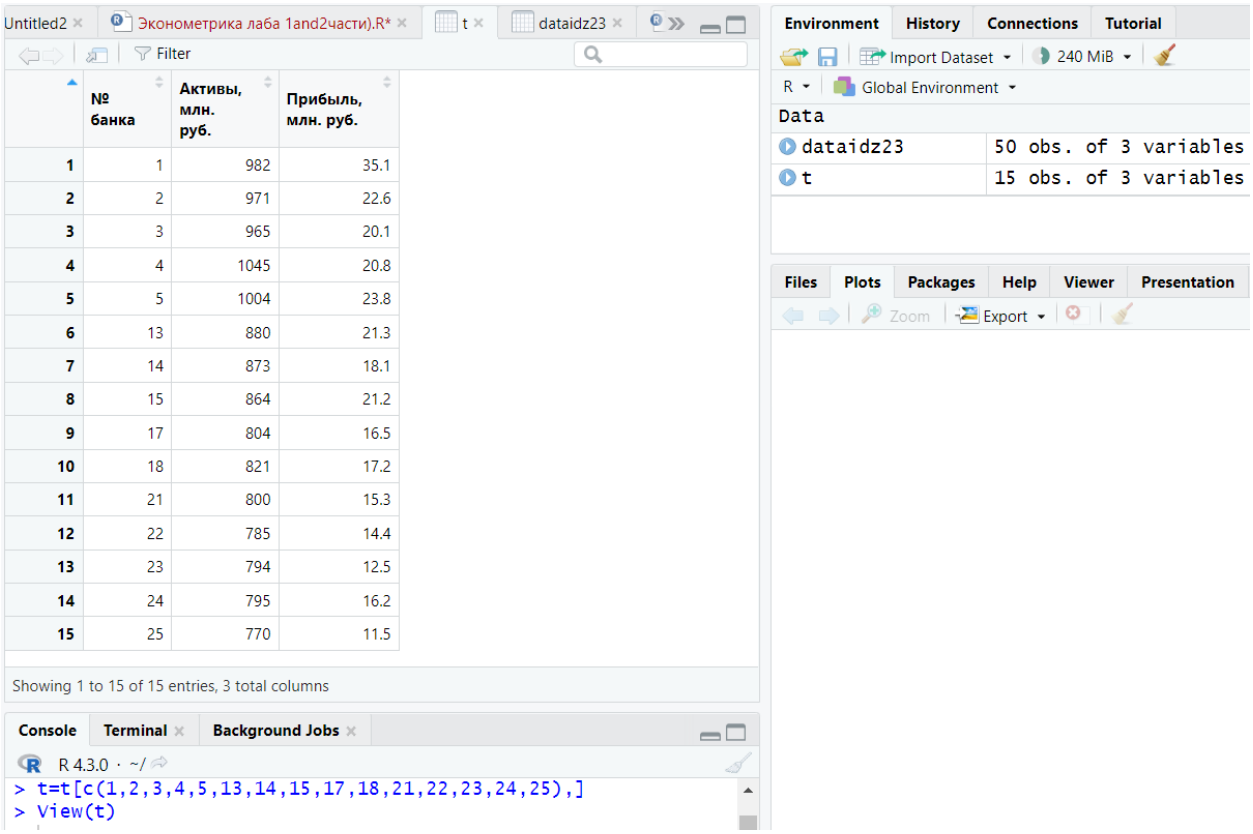

Import Excel Data
File/URL:
Browse...
Data Preview:
Choose File
2023 › Эконометрика › лабы
Поиск в: лабы
Упорядочить
Новая папка
ТАТЬЯНА — Личное
Рабочий стол
Загрузки
Документы
Изображения
Музыка
Видео
Лабораторные занятия
Имя
Дата изменения
dataidz
03.05.2023 10:39
dataidz23
08.09.2023 10:52
labafactor.R
02.05.2023 17:10
labafactor2.R
11.05.2023 13:30
ИД32Множественнаярегрессия-нов
02.05.2023 15:50
КТ эконометрика 2023
26.06.2023 21:57
Описание лабораторной работы №1
08.09.2023 10:52
Имя файла: dataidz23
All Files (*.*)
Open
Отмена
Cancel


Import Excel Data
File/URL:
~/учебная/2023/Эконометрика/лабы/dataidz23.xlsx
Browse...
Data Preview:
№ банка (double)
Активы, млн. руб. (double)
Прибыль, млн. руб. (double)
1 982 35.1
2 971 22.6
3 965 20.1
4 1045 20.8
5 1004 23.8
Previewing first 50 entries.
Import Options:
Name: dataidz23
Sheet: Default
Range: A1:D10
Max Rows:
Skip: 0
NA:
First Row as Names
Open Data Viewer
Code Preview:
library(readxl)
dataidz23 <- read_excel("учебная/2023/Эконометрика/лабы/dataidz23.xlsx")
View(dataidz23)
Reading Excel files using readxl
Import
Cancel

Результат:

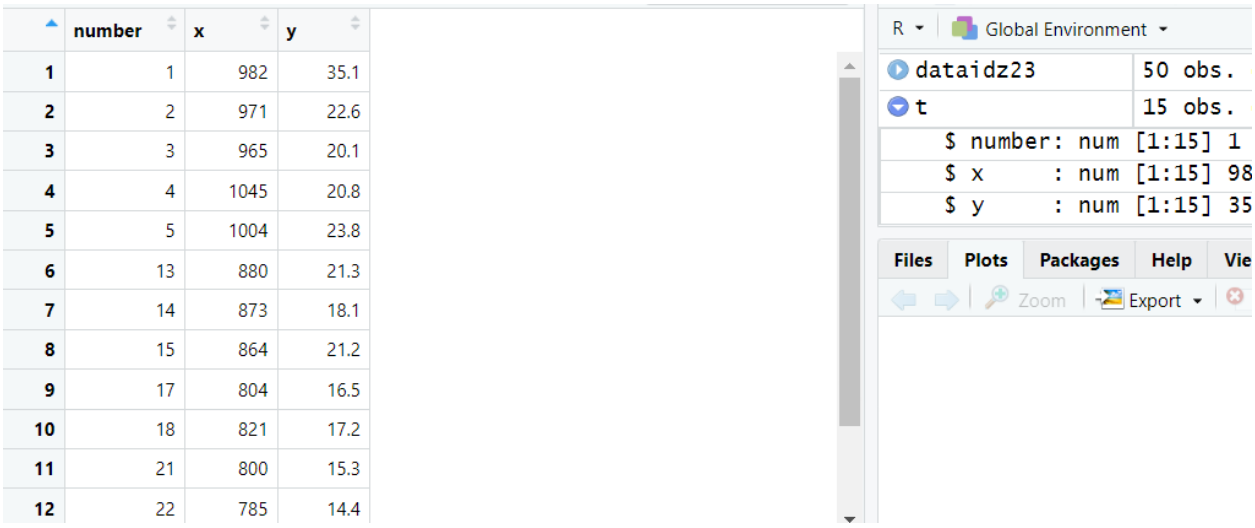

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled2
Эконометрика лаба 1and2части).R*
dataidz23
пример
№ банка
Активы, млн. руб.
Прибыль, млн. руб.
1 1 982 35.1
2 2 971 22.6
3 3 965 20.1
4 4 1045 20.8
5 5 1004 23.8
6 6 958 19.3
7 7 932 21.3
8 8 931 18.4
Showing 1 to 9 of 50 entries, 3 total columns
Environment History Connections Tutorial
Import Dataset 240 MiB
Global Environment
Data
dataidz23 50 obs. of 3 variables
Files Plots Packages Help Viewer Presentation
Export
Console Terminal Background Jobs
R 4.3.0 · ~/
> library(readxl)
> dataidz23 <- read_excel("учебная/2023/Эконометрика/лабы/dataidz23.xlsx")
> View(dataidz23)

Шаг 2. Формируем таблицу для своего варианта
#Чтобы не писать в названии переменных длинное название таблицы выполняем команду присвоения, выполнится команда если ее выделить мышью и нажать клавиши ctrl+enter:
t=dataidz23
#делаем свой вариант, сохраняя таблицу в том же объекте данных:
t=t[c(1,2,3,4,5,13,14,15,17,18,21,22,23,24,25),] перечень чисел дан в файле на втором листе dataidz23.



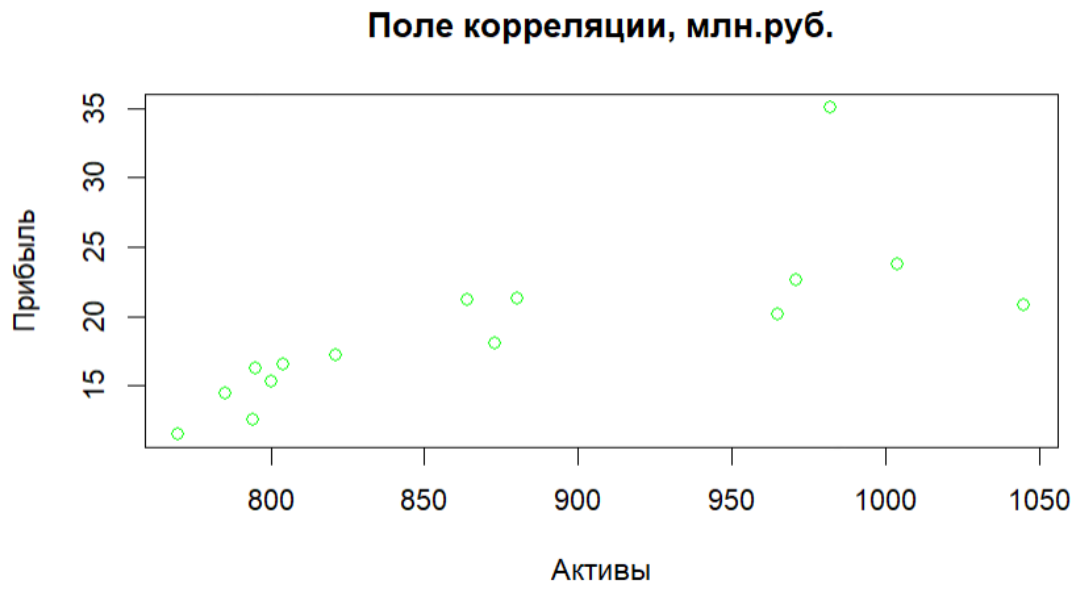
#переименуем в таблице названия столбцов, так как они указываются в названии переменных при обращении к ним в R:

names(t) =c("number", "x", "y")



Шаг. 3 Рисуем диаграмму рассеивания

plot(tx,ty, col='green', main="Поле корреляции, млн. руб.", xlim=c(min(t$x),max(t$x)), ylim=c(min(t$y),max(t$y)), xlab="Активы", ylab="Прибыль")



Вывод: связь линейная, но масштаб банка влияет на параметры модели.

Шаг 4. Введем новую переменную, которая будет указывать на то, крупный или мелкий банк:

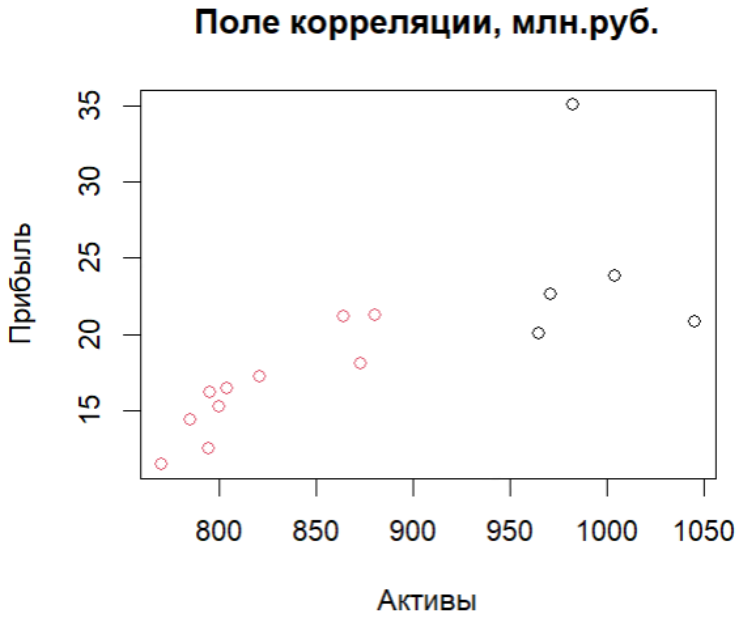
```
for (i in 1:15){if
      (t[i,2]< 900){
            t[i,4]=1}
      else {t[i,4]=0
      }
      i=i+1}
colnames(t)[4]="factor"
t$factor=as.factor(t$factor)
levels(t$factor)=c("high", "low")
```

Результат:

	number	x	y	factor
1	1	982	35.1	high
2	2	971	22.6	high
3	3	965	20.1	high
4	4	1045	20.8	high
5	5	1004	23.8	high
6	13	880	21.3	low
7	14	873	18.1	low
8	15	864	21.2	low
9	17	804	16.5	low
10	18	821	17.2	low
11	21	800	15.3	low
12	22	785	14.4	low
13	23	794	12.5	low
14	24	795	16.2	low
15	25	770	11.5	low

Построим график с учетом фактора:

plot(tx,ty, col = factor(t$factor), main="Поле корреляции, млн.руб.",
xlim=c(min(t$x),max(t$x)), ylim=c(min(t$y),max(t$y)), xlab="Активы", ylab="Прибыль")



Шаг. 5 Строим модели парной регрессии
5.1. Без учета масштаба банка
fit=lm(t$y~t$x)
summary(fit)

Результат:

```
41  fit=lm(t$y~t$x)
42  summary(fit)
43
44
45
```

46:1 (Top Level)

Console Terminal Background Jobs

R 4.3.0 · ~/

```
> summary(fit)

Call:
lm(formula = t$y ~ t$x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1241 -1.9278 -0.4353  0.8405 11.1051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.66382   10.01529  -2.163  0.04975 *
t$x           0.04650    0.01136   4.092  0.00127 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.937 on 13 degrees of freedom
Multiple R-squared:  0.5629,    Adjusted R-squared:  0.5293
F-statistic: 16.74 on 1 and 13 DF,  p-value: 0.001272
```

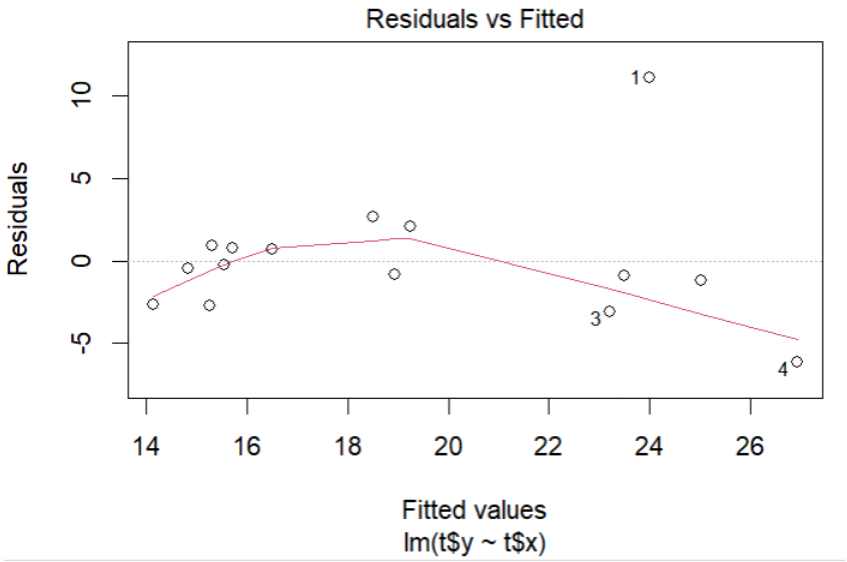
Выводы: Коэффициенты регрессии статистически значимо отличны от нуля (звездочки), но

```
Adjusted R-squared:  0.5293 < 0,75.
```

Проведем проверку выполнения предпосылок теоремы Гаусса-Маркова с помощью диагностических диаграмм:

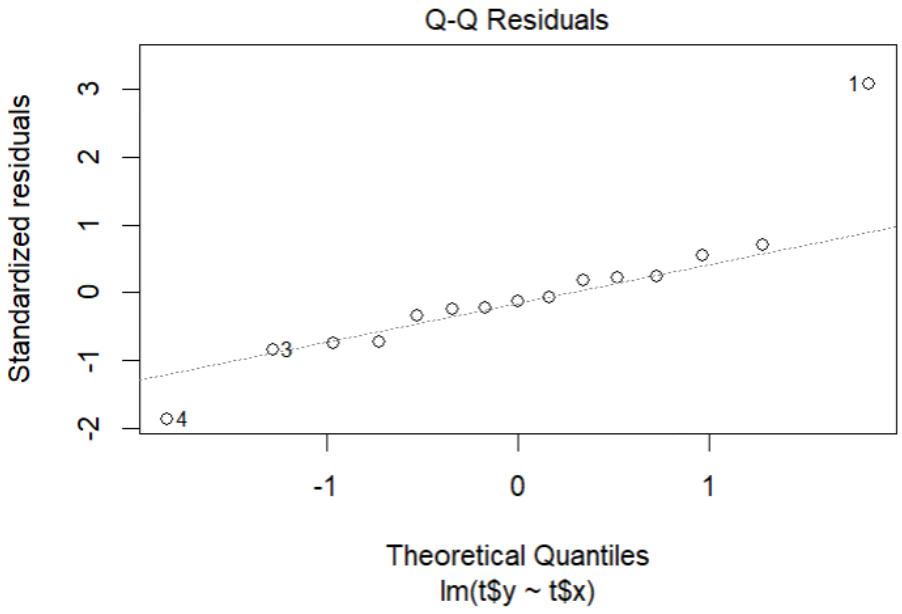
```
plot(fit)
```

1. Зависимость остатков от теоретических значений y

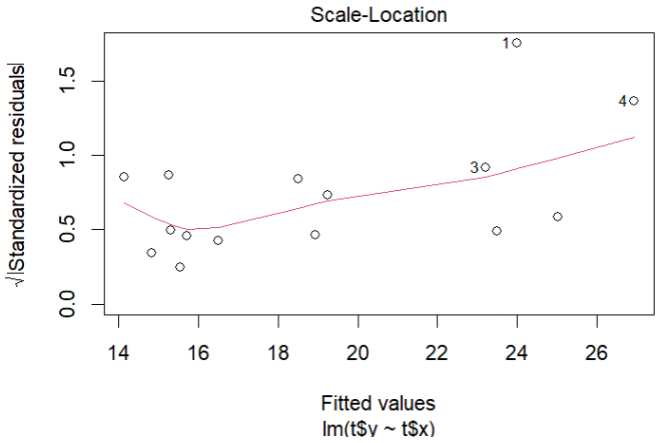


Вывод: наблюдение номер 4 и номер 1 нарушают линейность зависимости y от x.

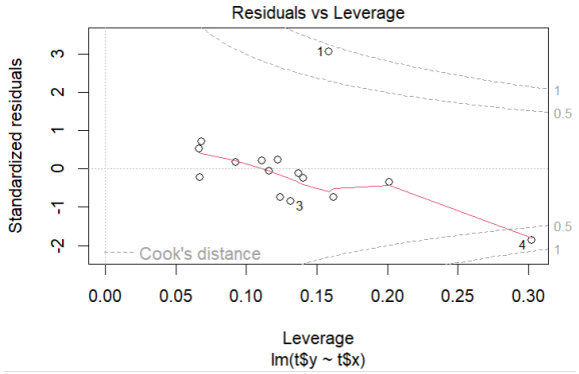
2. Q-Q график проверки согласованности остатков нормальному распределению:



Вывод: наблюдение номер 1 и номер 4 не позволяют считать остатки случайными.



3. Вывод: остатки имеют не постоянную дисперсию.

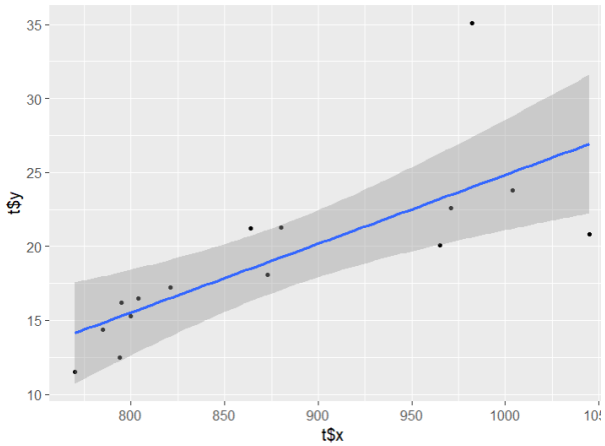


4. Наблюдения номер 4 и 1 влиятельные, то есть сильно отличаются от других. Итак, предпосылки теоремы нарушены.

Построим на графике линию регрессии с доверительным интервалом:

```
install.packages("ggplot2")

library(ggplot2)

ggplot(t, aes(t$x, t$y))+

    geom_point(size = 1)+

    geom_smooth(method = "lm")
```

Результат:



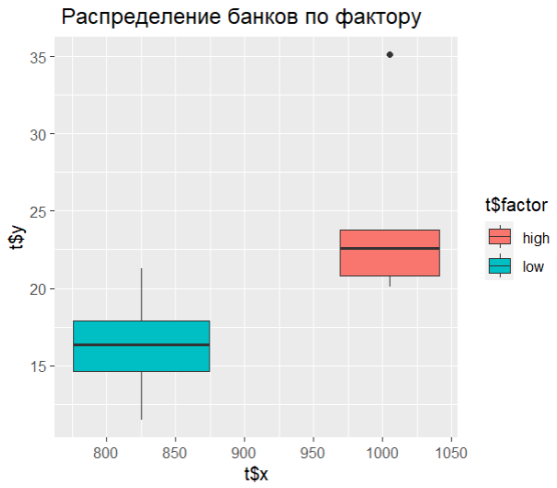
Видно, что есть 4 точки, которые не покрыты доверительным интервалом.

5.2. С учетом масштаба банка

1. Проверим данные на наличие выбросов:

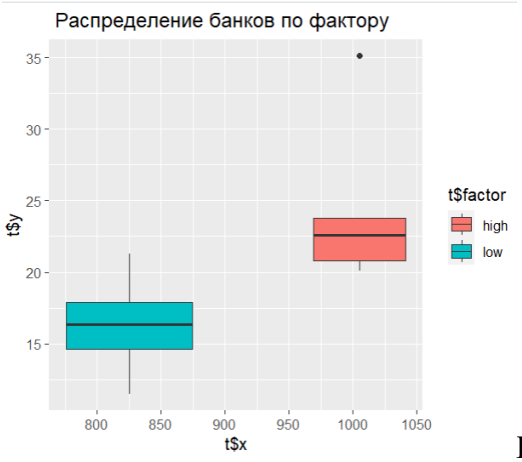
```
ggplot(t, aes(x = t$x, y = t$y, fill = t$factor))+
    geom_boxplot()+  ggtitle(" Распределение банков по фактору")
```

Результат:

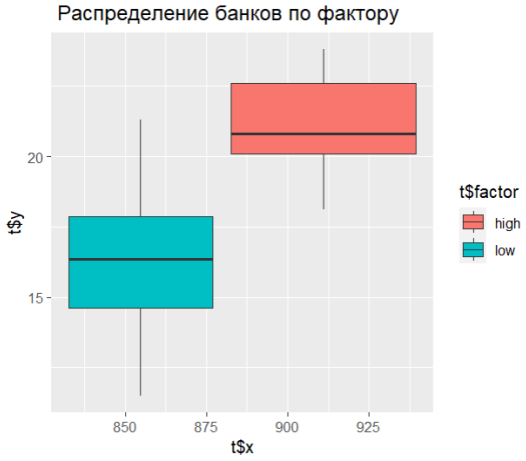


Видим, что в группе крупных банков есть выброс. Его нужно заменить на медиану.

```
t[4,2]=median(t$x)
```



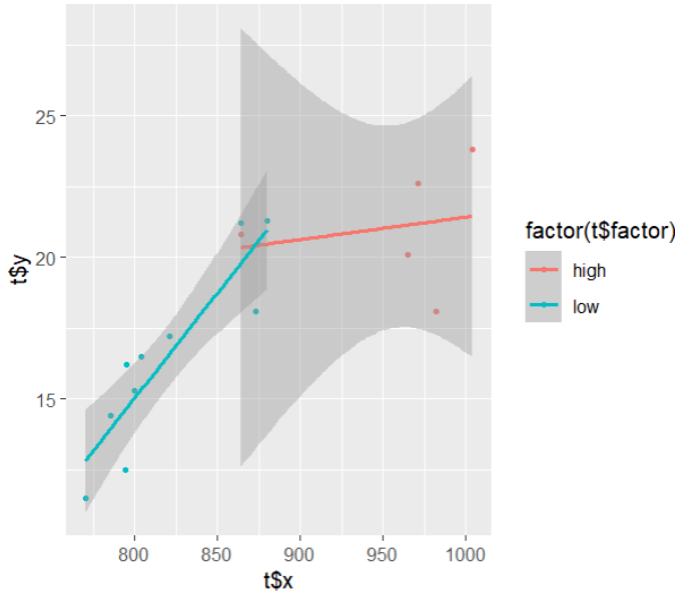
Выброс остался по y: t[1,3]=median(t$y)



Выбросов нет, видно, что группы отличаются по распределению признаков.

Строим регрессию с учетом фактора

ggplot(t, aes(tx, ty, col = factor(t$factor)))+ geom_point(size = 1)+ geom_smooth(method = "lm").



Проверим качество моделей:

Регрессия в группе мелких банков:

fit1=lm(t$y[t$factor=="low"]~ t$x[t$factor=="low"])

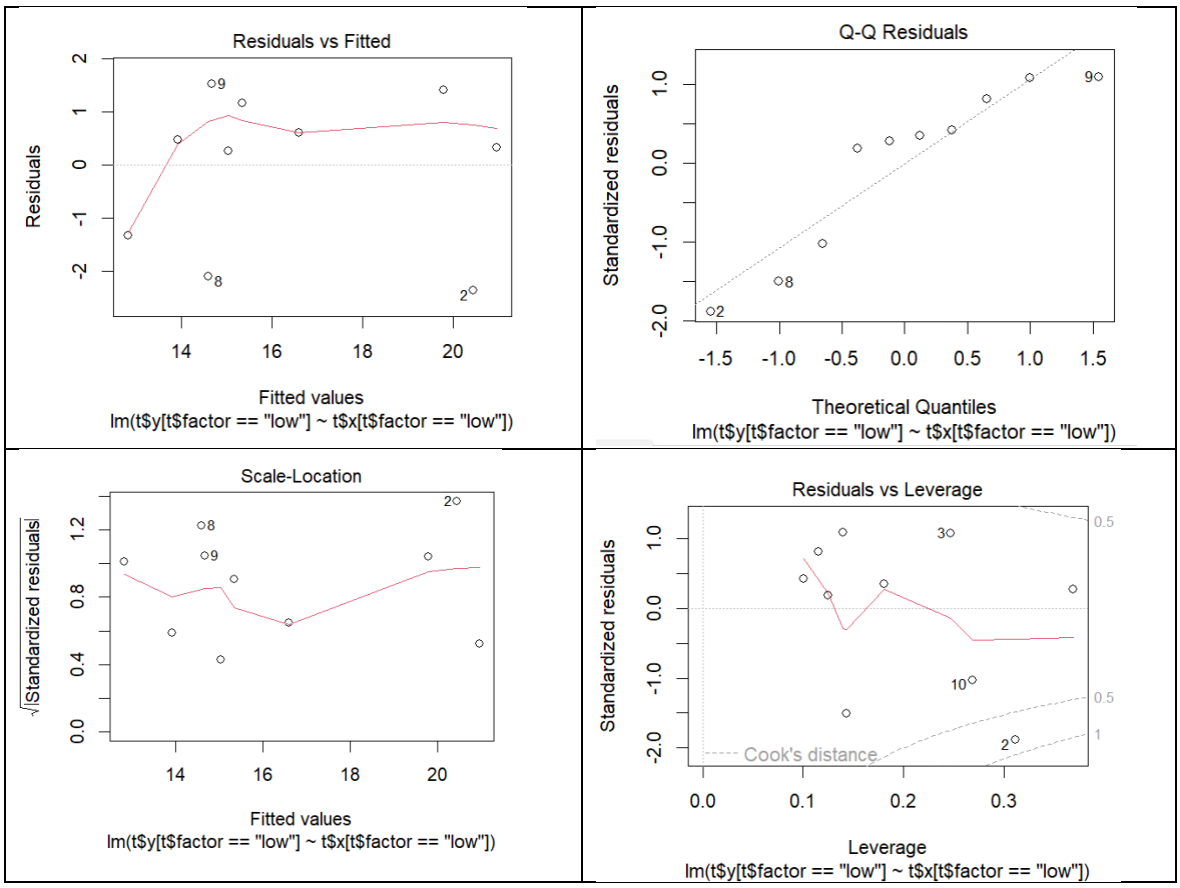
summary(fit1)

Результат:

```
Call:
lm(formula = t$y[t$factor == "low"] ~ t$x[t$factor == "low"])

Residuals:
    Min      1Q  Median      3Q     Max
-2.3516 -0.9240  0.3998  1.0220  1.5290

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -44.24747   10.41855  -4.247 0.002811 **
t$x[t$factor == "low"]    0.07411    0.01271   5.829 0.000392 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.505 on 8 degrees of freedom
Multiple R-squared:  0.8094,    Adjusted R-squared:  0.7856
F-statistic: 33.98 on 1 and 8 DF,  p-value: 0.000392
```

Вывод: качество улучшилось.

Проверим предпосылки:



Остатки случайны в группе мелких банков, а в целом по всей совокупности нет:

```
elow=fit$residuals
> shapiro.test(elow)

        Shapiro-Wilk normality test

data:  elow
W = 0.8497, p-value = 0.0172<0,05

> elow=fit1$residuals
> shapiro.test(elow)

        Shapiro-Wilk normality test

data:  elow
W = 0.86622, p-value = 0.0903>0,05
```

Проверим группу крупных банков:

fit2=lm(t$y[t$factor=="high"]~ t$x[t$factor=="high"])

summary(fit2)

ehigh=fit2$residuals

shapiro.test(ehigh)

Результаты:

```
Call:
lm(formula = t$y[t$factor == "high"] ~ t$x[t$factor == "high"])

Residuals:
      1       2       3       4       5
-3.1760  1.4109 -1.0417  0.4567  2.3501

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              13.513891  22.196151   0.609    0.586
t$x[t$factor == "high"]   0.007904   0.023159   0.341    0.755

Residual standard error: 2.51 on 3 degrees of freedom
Multiple R-squared:  0.03738,   Adjusted R-squared:  -0.2835
F-statistic: 0.1165 on 1 and 3 DF,  p-value: 0.7554

> ehigh=fit2$residuals
> shapiro.test(ehigh)

        Shapiro-Wilk normality test

data:  ehigh
W = 0.96265, p-value = 0.8263
```

Вывод: модель в группе крупных банков не является адекватной, хотя остатки случайные.

Задание для самостоятельной работы

Выполните примеры лабораторной работы на основе данных из табл. 1 и таблицы вариантов. Оформите отчёт.

Таблица 1

№ банка	Активы, млн. руб.	Прибыль, млн. руб.	№ банка	Активы, млн. руб.	Прибыль, млн. руб.
1	982	35,1	26	778	13,8
2	971	22,6	27	758	15,4
3	965	20,1	28	753	13,1
4	1045	20,8	29	720	12,5
5	1004	23,8	30	717	11,2
6	958	19,3	31	716	13,3
7	932	21,3	32	712	8,6
8	931	18,4	33	703	8,3
9	928	20,2	34	690	5,7
10	924	19,4	35	684	7,5
11	921	20,6	36	677	5,7
12	901	15,6	37	673	5,2
13	880	21,3	38	649	4,7
14	873	18,1	39	631	6,7
15	864	21,2	40	627	4,8
16	859	18,4	41	609	8,9

17	804	16,5	42	605	6,7
18	821	17,2	43	574	5,1
19	801	18	44	563	6,3
20	801	19,4	45	556	6,3
21	800	15,3	46	543	3,6
22	785	14,4	47	538	5,3
23	794	12,5	48	526	5
24	795	16,2	49	510	5,8
25	770	11,5	50	512	5,1

Таблица вариантов

№ варианта	Перечень номеров банков, данные о которых необходимо взять из массива исходных данных для выполнения своего варианта
1	4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 5, 11, 13, 1, 9, 11
2	1, 2. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 40, 45
3	14, 16, 17, 19, 20, 22, 23, 25, 26, 28, 29, 31, 32, 34, 35, 37, 38, 39, 40, 41, 43, 44, 46, 47, 49, 50
4	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 37, 38, 39, 40, 41,44
5	6, 8, 9, 14, 15, 18, 19, 21, 22, 23, 26, 27, 28, 30, 32, 34, 35, 37, 39, 41, 43, 44, 45, 47, 49, 50, 1
6	1, 3, 5, 9, 10, 12, 14, 18, 19, 21, 23, 25, 27, 29, 30, 31, 33, 34, 35, 36, 39, 40, 41, 42, 43, 44, 50
7	1, 2, 14, 15, 16, 20, 21, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 36, 38, 39, 40, 41, 42, 50, 49
8	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 48, 49, 50
9	3, 5, 6, 7, 9, 10, 11, 12, 14, 15, 16, 20, 21, 22, 26, 27, 28, 29, 34, 35, 36, 37, 38, 39, 40, 50, 49
10	10, 11, 13, 14, 15, 16, 17, 20, 21, 22, 23, 24, 26, 28, 30, 32, 35, 37, 38, 40, 41, 43, 46, 47, 48
11	1, 8, 12, 16, 20, 24, 28, 32, 36, 41, 44, 48, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 5, 11, 13, 49, 50
12	1, 2. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 34, 35, 36, 37
13	14, 16, 17, 19, 20, 22, 23, 25, 6, 28, 29, 31, 32, 34, 35, 37, 48, 39, 40, 41, 43, 44, 46, 47, 49, 50
14	7, 18, 19, 10, 11, 12, 13, 14, 15, 16, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 37, 38, 39, 40, 41
15	6, 8, 9, 14, 15, 18, 19, 21, 22, 23, 26, 27, 28, 30, 32, 34, 35, 37, 39, 41, 43, 44, 45, 47, 49, 50
16	1, 3, 5, 9, 10, 12, 14, 18, 19, 21, 23, 25, 27, 29, 30, 31, 33, 34, 35, 36, 39, 40, 41, 42, 43, 44
17	1, 2, 14, 15, 16, 20, 21, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 36, 38, 39, 50
18	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 48, 49, 50
19	3, 5, 6, 7, 9, 10, 11, 12, 14, 15, 16, 20, 21, 22, 26, 27, 28, 29, 34, 35, 36, 37
20	10, 11, 13, 14, 15, 16, 17, 20, 21, 22, 23, 24, 26, 28, 30, 32, 35, 37, 38, 40, 41, 43, 46, 47, 48, 49, 50
21	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41
22	3, 5, 6, 7, 9, 10, 11, 12, 14, 15, 16, 20, 21, 22, 26, 27, 28, 29, 34, 35, 36
23	10, 11, 13, 14, 15, 16, 17, 20, 21, 22, 23, 24, 26, 28, 30, 32, 35, 37, 38, 40, 41, 43, 46, 47, 48, 49, 50
24	1, 8, 12, 16, 20, 24, 28, 32, 36, 41, 44, 48, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 5, 11, 13, 50
25	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23

Примечание: если варианта не хватает, осуществите случайную выборку не менее 30 банков sample(x=data, size=30)