



Программирование в среде R

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

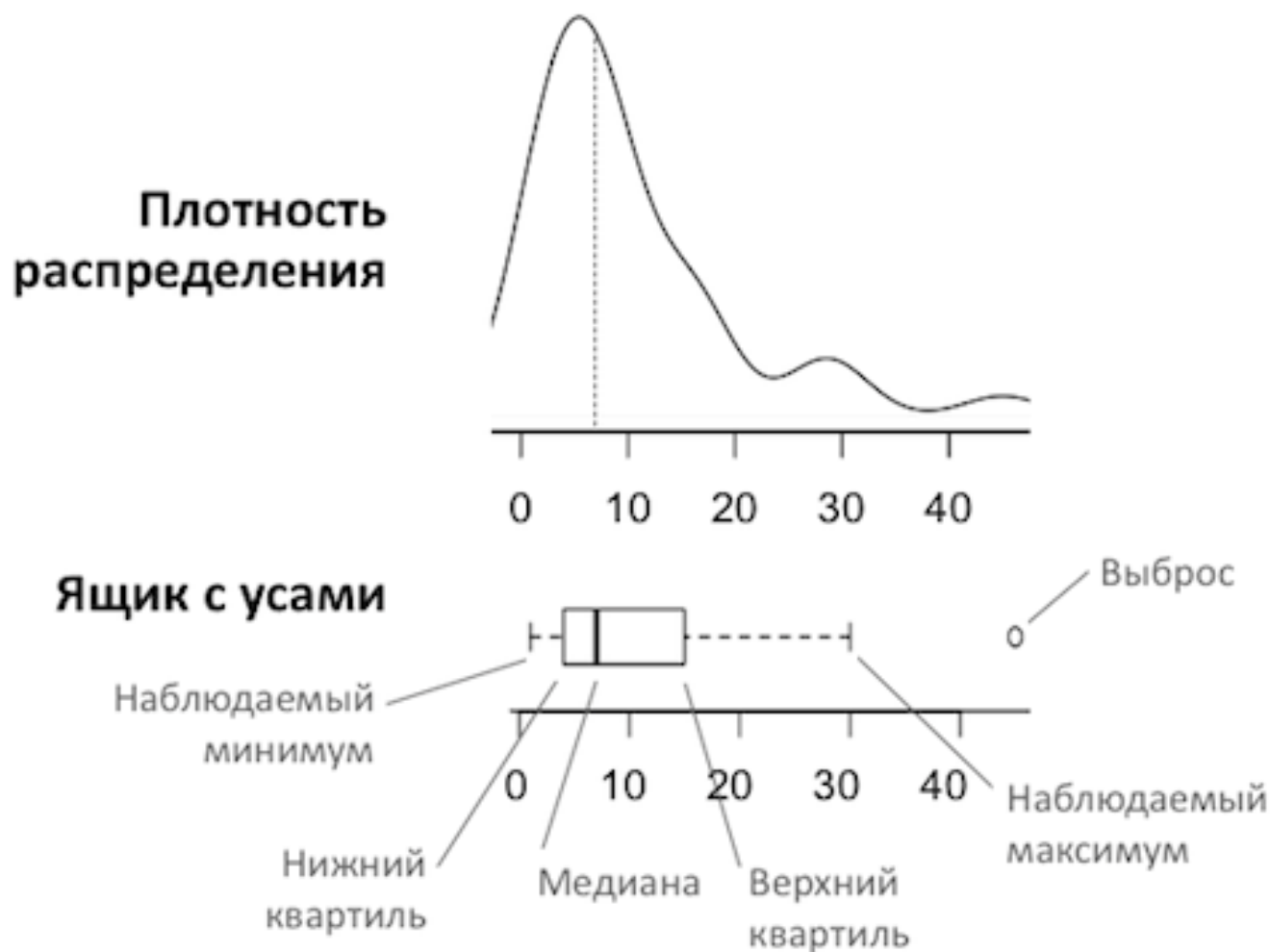
mpg	расход топлива (количества миль на галлон топлива)
cyl	кол-во цилиндров
disp	объем двигателя
hp	мощность двигателя (лошадиные силы)
drat	передаточное число заднего моста
wt	вес
qsec	значение времени разгона
vs	тип двигателя (v-образный, рядный)
am	тип коробки передач
gear	кол-во передач
carb	число карбюраторов

Диаграммы размахов

boxplot()

- Диаграммы размахов (box plot) иллюстрируют распределение значений непрерывной переменной, отображая пять параметров:
 - минимум,
 - нижний квартиль (25-й процентиль),
 - медиану (50-й процентиль),
 - верхний квартиль (75-й процентиль)
 - максимум.
- На этой диаграмме также могут быть отображены вероятные выбросы (значения, выходящие за диапазон в ± 1.5 межквартильного размаха, разности верхней и нижней квартилей).
- По умолчанию каждый «ус» продолжается до минимального или максимального значения, которое не выходит за пределы 1.5 межквартильного размаха. Выходящие за эти пределы значения отмечаются точками

Сравнение плотности распределения и ящика с усами

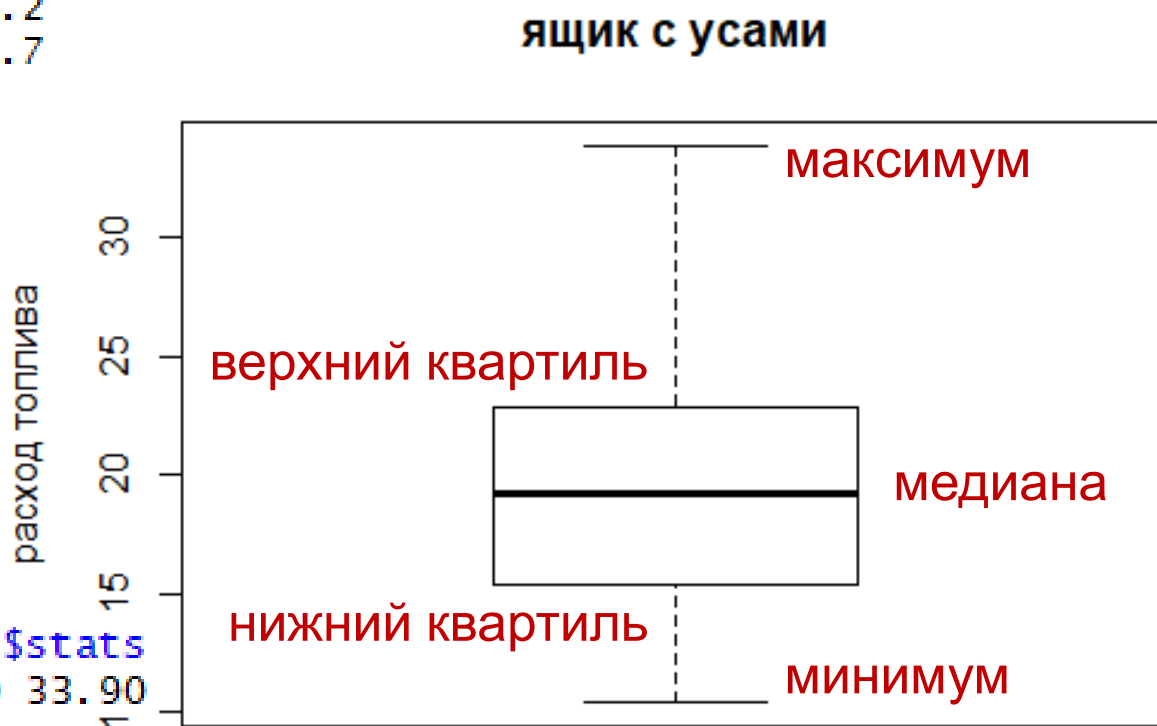


boxplot()

```
boxplot(mtcars$mpg, main="ящик с усами", ylab="расход топлива")
```

```
> mtcars$mpg  
[1] 21.0 21.0 22.8 21.4 18.7  
[6] 18.1 14.3 24.4 22.8 19.2  
[11] 17.8 16.4 17.3 15.2 10.4  
[16] 10.4 14.7 32.4 30.4 33.9  
[21] 21.5 15.5 15.2 13.3 19.2  
[26] 27.3 26.0 30.4 15.8 19.7  
[31] 15.0 21.4
```

```
> boxplot.stats(mtcars$mpg)$stats  
[1] 10.40 15.35 19.20 22.80 33.90
```



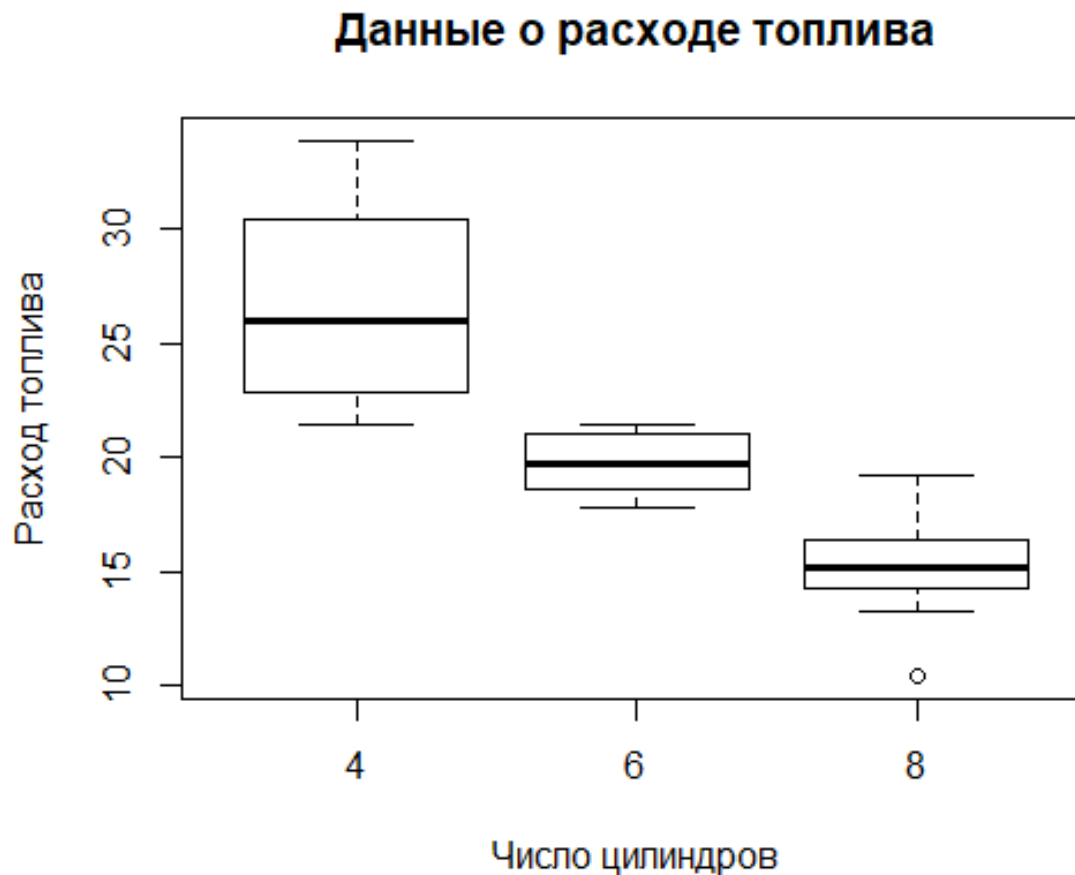
Использование диаграмм размахов для сравнения групп между собой

- Диаграммы размахов можно построить для отдельных переменных или для групп переменных.
- Общий вид команды таков: `boxplot(formula, data=dataframe)`
- где *formula* – это формула, а *dataframe* обозначает таблицу данных (или список), где содержатся данные.
- Примером формулы может служить выражение $y \sim A$, где для каждого значения категориальной переменной A будет построена отдельная диаграмма размахов для числовой переменной y . Формула $y \sim A * B$ позволит получить отдельные диаграммы размахов для всех комбинаций значений переменной y , заданных категориальными переменными A и B .

Использование диаграмм размахов для сравнения групп между собой

```
boxplot(mpg ~ cyl, data=mtcars, main="Данные о расходе топлива",  
xlab="Число цилиндров", ylab="Расход топлива")
```

*~ способ
записи формул,
описывающих
связь между
переменными*



Параметры

```
boxplot(mpg ~ cyl, data=mtcars,  
varwidth=TRUE, main="Данные о  
расходе топлива", xlab="Число  
цилиндров", ylab="Расход топлива")
```

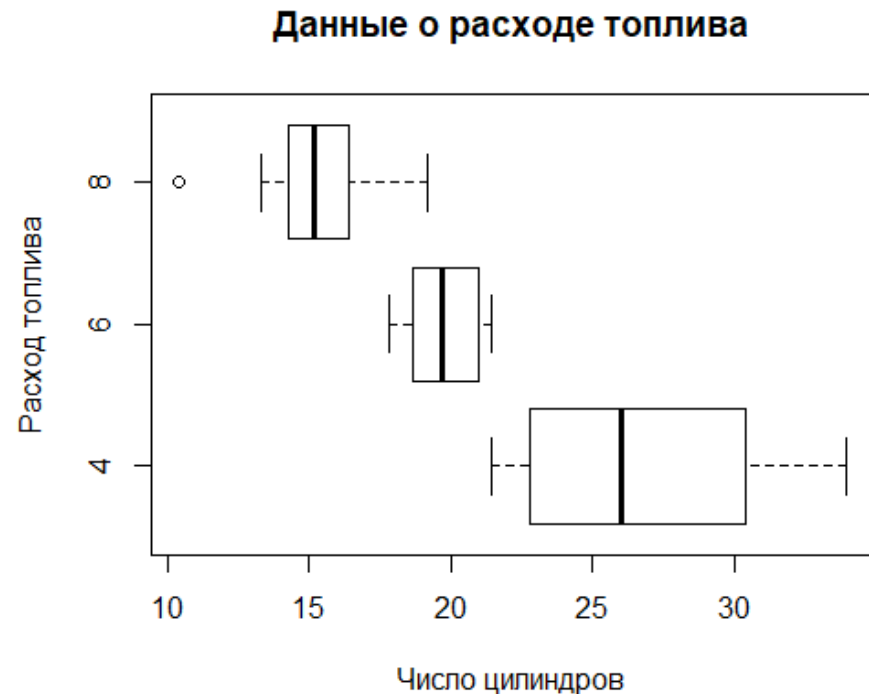
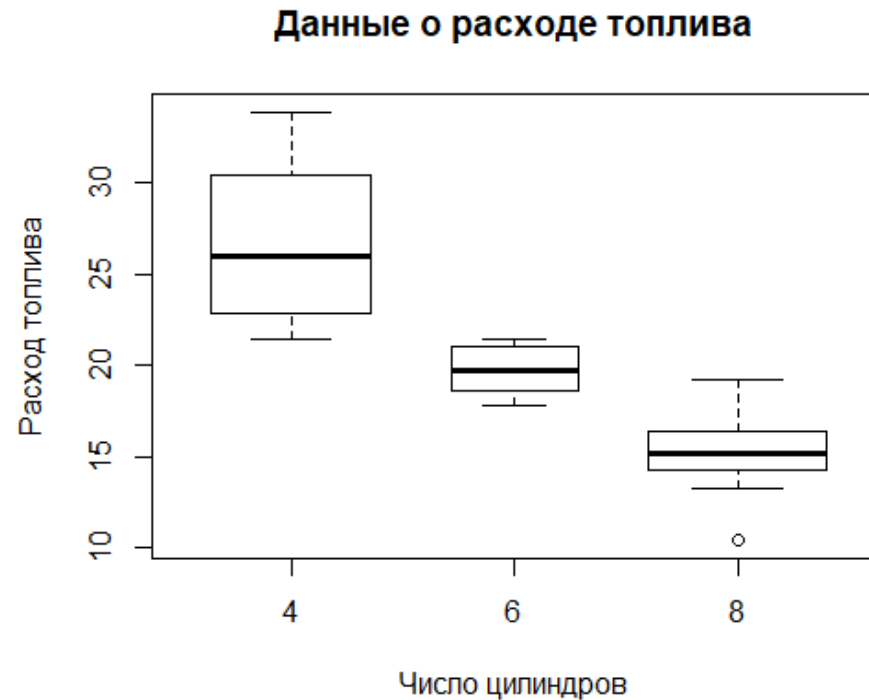
varwidth=TRUE

ширина "ящиков" будет
пропорциональна квадратному
корню из размера выборки

```
boxplot(mpg ~ cyl, data=mtcars,  
horizontal=TRUE, main="Данные о  
расходе топлива", xlab="Число  
цилиндров", ylab="Расход топлива")
```

horizontal=TRUE

поменять оси местами

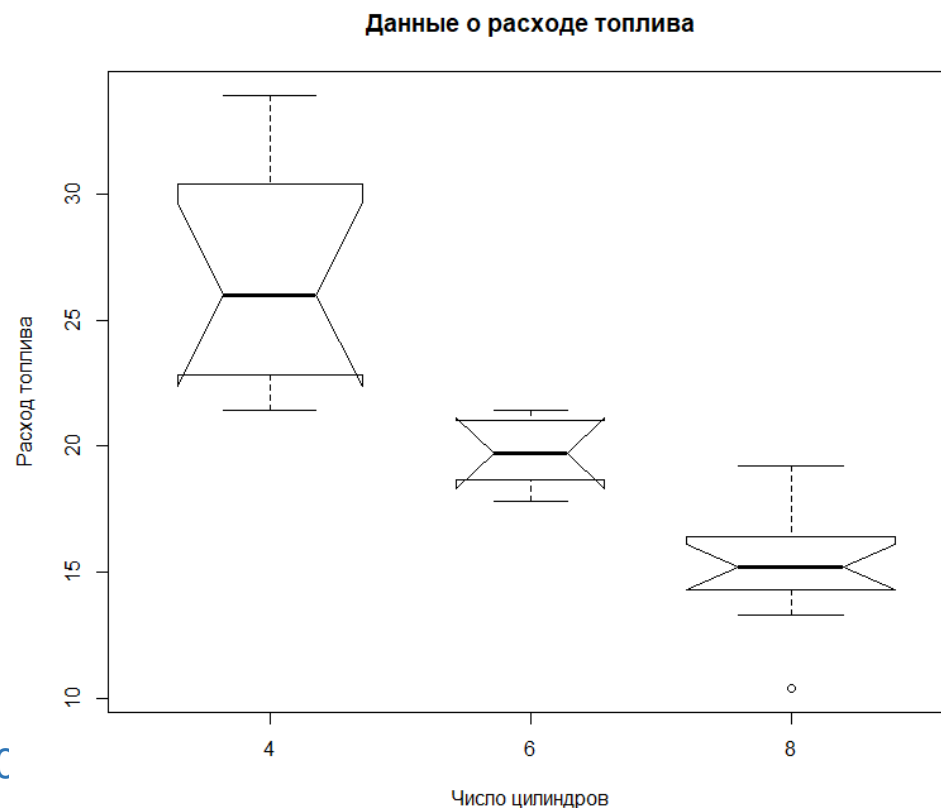


Параметры

notch=TRUE

получатся "ящики" с "насечками".
Если "насечки" двух ящиков не
перекрываются, высока
вероятность того, что медианы
соответствующих совокупностей
различаются

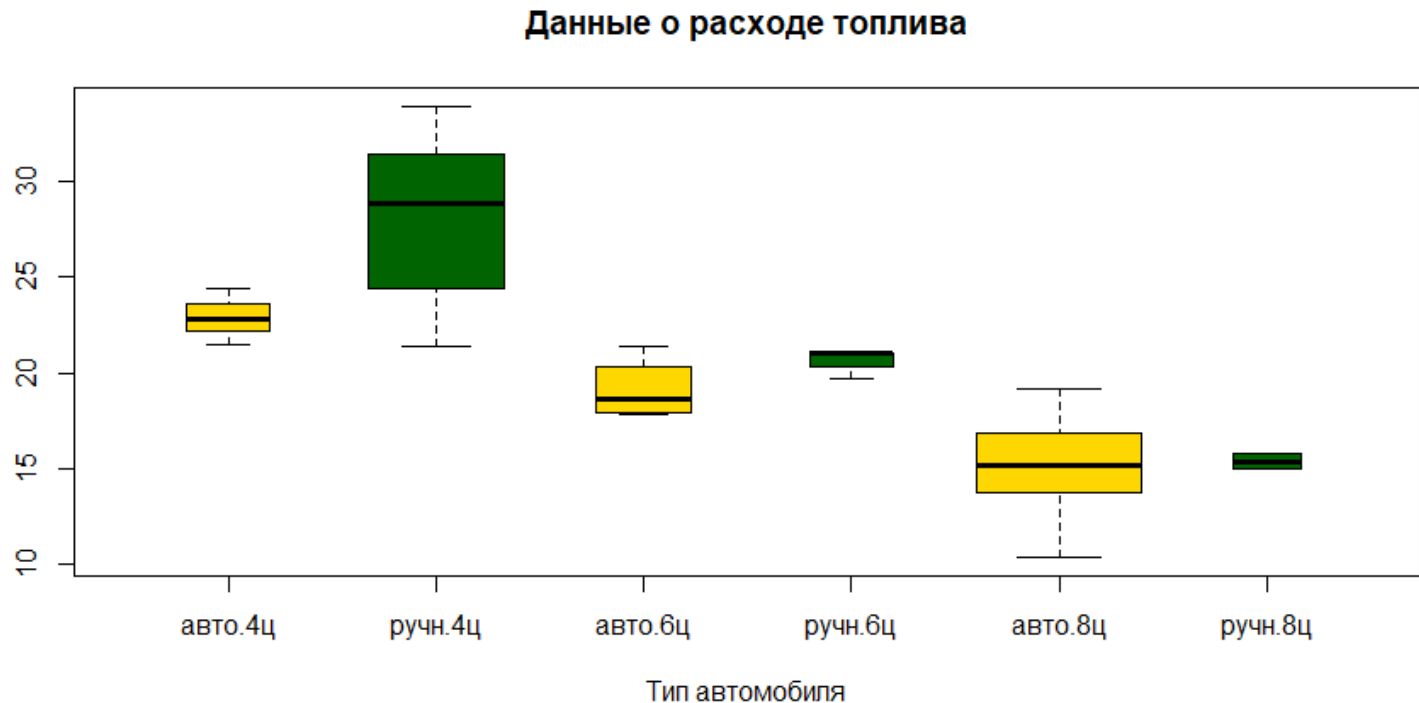
```
boxplot(mpg ~ cyl, data=mtcars,  
notch=TRUE, varwidth=TRUE,  
main="Данные о расходе  
топлива", xlab="Число  
цилиндров", ylab="Расход  
топлива")
```



Диаграммы размахов для нескольких группирующих переменных

Формула $y \sim A * B$

```
mtcars$cyl.f <- factor(mtcars$cyl, levels=c(4,6,8), labels=c("4ц", "6ц", "8ц"))  
mtcars$am.f <- factor(mtcars$am, levels=c(0,1), labels=c("авто", "ручн"))  
boxplot(mpg ~ am.f * cyl.f, data=mtcars, varwidth=TRUE,  
col=c("gold", "darkgreen"), main="Данные о расходе топлива",  
xlab="Тип автомобиля")
```



Скрипичные диаграммы

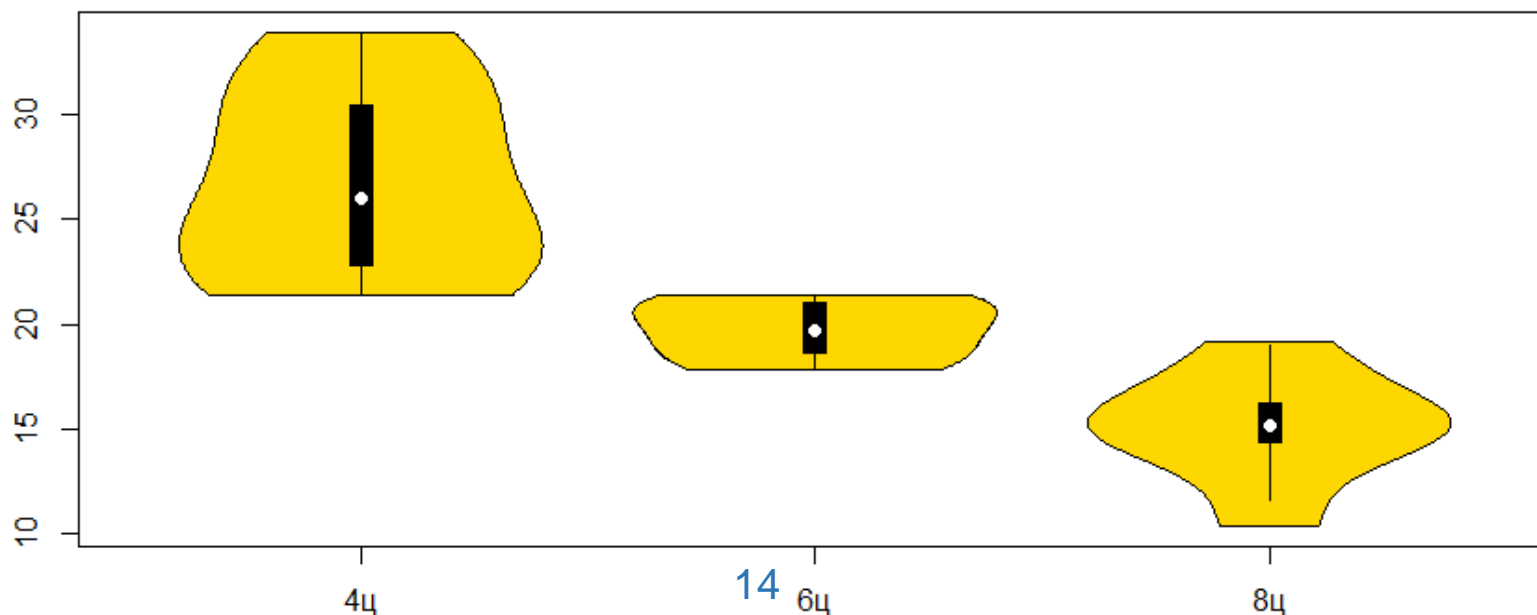
Скрипичные диаграммы

- Скрипичные диаграммы – модификация диаграмм размахов, сочетание диаграммы размахов и диаграммы ядерной оценки функции плотности.
- Создается при помощи функции `vioplot()` из пакета `vioplot`
- `vioplot(x1, x2, ... , names=, col=)`
 - где `x1, x2, ...` – это один или более числовых векторов, которые нужно изобразить графически (для каждого вектора будет построена своя скрипичная диаграмма).
 - Параметр `names` задает текстовый вектор с подписями для диаграмм,
 - `col` – вектор, содержащий названия цветов каждой диаграммы

vioplot

```
install.packages("vioplot")  
library(vioplot)  
x1 <- mtcars$mpg[mtcars$cyl==4]  
x2 <- mtcars$mpg[mtcars$cyl==6]  
x3 <- mtcars$mpg[mtcars$cyl==8]  
vioplot(x1, x2, x3, names=c("4ц", "6ц", "8ц"), col="gold")  
title("Скрипичные диаграммы расхода топлива")
```

Скрипичные диаграммы расхода топлива



- Скрипичные диаграммы представляют собой симметричные диаграммы ядерной оценки функции плотности, наложенные на диаграммы размахов.
 - белая точка – медиана
 - черный прямоугольник – межквартильный размах
 - тонкие черные линии – "усы"
 - Внешний контур фигуры – это диаграмма ядерной оценки функции плотности.

Точечные диаграммы

- `dotchart(x, labels=)`
- где x – это числовой вектор,
- `labels` задает вектор, в котором содержатся подписи к каждой точке.
- параметр `groups` назначает фактор, определяющего группировку элементов вектора x .
- параметр `gcolor` определяет цвет подписей для разных групп
- параметр `sx` определяет размер подписей.

dotchart()

```
x <- mtcars[order(mtcars$mpg),]  
x$cyl <- factor(x$cyl)  
x$color[x$cyl==4] <- "red"  
x$color[x$cyl==6] <- "blue"  
x$color[x$cyl==8] <- "darkgreen"  
dotchart(x$mpg, labels = row.names(x), cex=.7, groups = x$cyl, gcolor =  
"black", color = x$color, pch=19, main = "Расход топлива, группировка по  
числу цилиндров", xlab = "Миль на галлон")
```



Корреляции

Коэффициенты корреляции

- Коэффициенты корреляции используются для описания связей между количественными переменными.
- Знак коэффициента (+ или –) свидетельствует о направлении связи (положительная или отрицательная)
- Величина коэффициента показывает силу связи (варьирует от 0 – нет связи до 1 – абсолютно предсказуемая взаимосвязь).

cor()

- `cor(x, use= , method=)`
- Линейный коэффициент корреляции Пирсона (Pearson product moment correlation) отражает степень линейной связи между двумя количественными переменными.
- Коэффициент ранговой корреляции Спирмана (Spearman's Rank Order correlation) – мера связи между двумя ранжированными переменными.
- Коэффициент Тау Кэнделла (Kendall's Tau) – также непараметрический показатель ранговой корреляции.

cor(). Параметры

- `x` - Матрица или таблица данных
- `use`. Упрощает работу с пропущенными данными.
 - `all.obs` (предполагается, что пропущенные значения отсутствуют; их наличие вызовет сообщение об ошибке),
 - `everything` (любая корреляция, включающая строку с пропущенным значением, не будет вычисляться – обозначается как `missing`),
 - `complete.obs` (учитываются только строки без пропущенных значений)
 - `pairwise.complete.obs` (учитываются все полные наблюдения для каждой пары переменных в отдельности)
- `method`. Определяет тип коэффициента корреляции. Возможные значения
 - `pearson` (по умолчанию)
 - `spearman`
 - `kendall`

cor(mtcars)

```
> mtcars
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3

```
> t1 <- mtcars[,c(1,2,9)]
```

```
> cor(t1)
```

	mpg	cyl	am
mpg	1.0000000	-0.852162	0.5998324
cyl	-0.8521620	1.0000000	-0.5226070
am	0.5998324	-0.522607	1.0000000

cor(mtcars, method=)

```
> t1 <- mtcars[,c(1,2,9)]
> cor(t1,method="pearson")
              mpg          cyl          am
mpg  1.0000000 -0.852162  0.5998324
cyl -0.8521620  1.000000 -0.5226070
am   0.5998324 -0.522607  1.0000000
> cor(t1,method="spearman")
              mpg          cyl          am
mpg  1.0000000 -0.9108013  0.5620057
cyl -0.9108013  1.0000000 -0.5220712
am   0.5620057 -0.5220712  1.0000000
> cor(t1,method="kendall")
              mpg          cyl          am
mpg  1.0000000 -0.7953134  0.4690128
cyl -0.7953134  1.0000000 -0.4946212
am   0.4690128 -0.4946212  1.0000000
```

Ковариация

```
> cov(t1)
```

	mpg	cyl	am
mpg	36.324103	-9.1723790	1.8039315
cyl	-9.172379	3.1895161	-0.4657258
am	1.803931	-0.4657258	0.2489919

Результат – прямоугольная матрица

```
> t1 <- mtcars[,c(1,2,9)]  
> t2 <- mtcars[,c(1,2)]  
> cor(t1,t2,method="pearson")
```

	mpg	cyl
mpg	1.0000000	-0.852162
cyl	-0.8521620	1.000000
am	0.5998324	-0.522607

Частные корреляции

- Частная корреляция – это корреляция между двумя количественными переменными, зависящими, в свою очередь, от одной или более других количественных переменных.
- Для вычисления коэффициентов частной корреляции можно использовать функцию `pcor()` из пакета `ggm`.
- `pcor(u, S)`
 - u – это числовой вектор, в котором первые два числа – это номера переменных, для которых нужно вычислить коэффициент, а остальные числа – номера «влияющих» переменных (воздействие которых должно быть отделено)
 - S – это ковариационная матрица для всех этих переменных.

Частные корреляции

```
> pcor(c(1,2,3,4,5,6,7,8,9,10,11),cov(mtcars))  
[1] -0.02326429  
> pcor(c(1,2,3,4,5,6,7,8,10,11),cov(mtcars))  
[1] -0.0926765  
> pcor(c(1,2,3,4,5,6,7,8,10),cov(mtcars))  
[1] -0.1127779
```

Проверка статистической значимости корреляций

- Стандартная нулевая гипотеза – это отсутствие связи (то есть коэффициент корреляции для генеральной совокупности равен нулю).
- Для проверки значимости отдельных корреляционных коэффициентов Пирсона, Спирмена и Кэнделла можно использовать функцию `cor.test()`.
- `cor.test(x, y, alternative = , method =)`
 - где x и y – это переменные, корреляция между которыми исследуется,
 - опция `alternative` определяет тип теста (“two.side”, “less” или “greater”),
 - опция `method` задает тип корреляции (“pearson”, “kendall” или “spearman”).
 - опция `alternative="less"` для проверки гипотезы о том, что в генеральной совокупности коэффициент корреляции меньше нуля
 - опция `alternative="greater"` – для проверки того, что он больше нуля. По умолчанию `alternative="two.side"` (проверяется гипотеза о том, что коэффициент корреляции в генеральной совокупности не равен нулю).

cor.test()

```
> cor.test(mtcars$mpg,mtcars$cyl)

Pearson's product-moment correlation

data:  mtcars$mpg and mtcars$cyl
t = -8.9197, df = 30, p-value = 6.113e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9257694 -0.7163171
sample estimates:
      cor 
-0.852162
```

Нулевая гипотеза заключается в том, что коэффициент корреляции Пирсона между расходом топлива и количеством цилиндров равен нулю. Если этот коэффициент для генеральной совокупности равен нулю, то его значение для случайной выборки будет равно 0.852 реже, чем в одном случае из 10^{10} ($p\text{-value} = 6.113e - 10$).

Учитывая, насколько мала вероятность, мы отвергнем нулевую гипотезу и примем альтернативную – о том, что значение этого коэффициента для генеральной совокупности *не* равно нулю.

corr.test()

- При помощи функции `cor.test()` одновременно можно проверить значимость только одного коэффициента корреляции.
- В пакете `psych` есть функция `corr.test()`, которая позволяет вычислить коэффициенты корреляции Пирсона, Спирмена и Кэнделла между несколькими переменными и оценить их достоверность.
- `corr.test(x, use =, method=)`
 - `use=` может принимать значения "pairwise" или "complete" (для попарного или построчного удаления пропущенных значений соответственно).
 - Значения опции `method=` "pearson" (по умолчанию), "spearman" или "kendall".

corr.test(mtcars, method="pearson")

```
> corr.test(mtcars, method="pearson")
Call:corr.test(x = mtcars, method = "pearson")
Correlation matrix
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.60	0.48	-0.55
cyl	-0.85	1.00	0.90	0.83	-0.70	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.90	1.00	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1.00	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.70	-0.71	-0.45	1.00	-0.71	0.09	0.44	0.71	0.70	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1.00	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00	0.17	0.21	-0.57
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00

```
Sample size
[1] 32
Probability values (Entries above the diagonal are adjusted for multiple tests.)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	0.00	0	0.00	0.00	0.00	0.00	0.22	0.00	0.01	0.10	0.02
cyl	0.00	0	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.08	0.04
disp	0.00	0	0.00	0.00	0.00	0.00	0.20	0.00	0.01	0.02	0.30
hp	0.00	0	0.00	0.00	0.17	0.00	0.00	0.00	1.00	1.00	0.00
drat	0.00	0	0.00	0.01	0.00	0.00	1.00	0.19	0.00	0.00	1.00
wt	0.00	0	0.00	0.00	0.00	0.00	1.00	0.02	0.00	0.01	0.20
qsec	0.02	0	0.01	0.00	0.62	0.34	0.00	0.00	1.00	1.00	0.00
vs	0.00	0	0.00	0.00	0.01	0.00	0.00	0.00	1.00	1.00	0.02
am	0.00	0	0.00	0.18	0.00	0.00	0.21	0.36	0.00	0.00	1.00
gear	0.01	0	0.00	0.49	0.00	0.00	0.24	0.26	0.00	0.00	1.00
carb	0.00	0	0.03	0.00	0.62	0.01	0.00	0.00	0.75	0.13	0.00

corr.test(mtcars, method="spearman")

```
> corr.test(mtcars, method="spearman")
Call:corr.test(x = mtcars, method = "spearman")
Correlation matrix
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.91	-0.91	-0.89	0.65	-0.89	0.47	0.71	0.56	0.54	-0.66
cyl	-0.91	1.00	0.93	0.90	-0.68	0.86	-0.57	-0.81	-0.52	-0.56	0.58
disp	-0.91	0.93	1.00	0.85	-0.68	0.90	-0.46	-0.72	-0.62	-0.59	0.54
hp	-0.89	0.90	0.85	1.00	-0.52	0.77	-0.67	-0.75	-0.36	-0.33	0.73
drat	0.65	-0.68	-0.68	-0.52	1.00	-0.75	0.09	0.45	0.69	0.74	-0.13
wt	-0.89	0.86	0.90	0.77	-0.75	1.00	-0.23	-0.59	-0.74	-0.68	0.50
qsec	0.47	-0.57	-0.46	-0.67	0.09	-0.23	1.00	0.79	-0.20	-0.15	-0.66
vs	0.71	-0.81	-0.72	-0.75	0.45	-0.59	0.79	1.00	0.17	0.28	-0.63
am	0.56	-0.52	-0.62	-0.36	0.69	-0.74	-0.20	0.17	1.00	0.81	-0.06
gear	0.54	-0.56	-0.59	-0.33	0.74	-0.68	-0.15	0.28	0.81	1.00	0.11
carb	-0.66	0.58	0.54	0.73	-0.13	0.50	-0.66	-0.63	-0.06	0.11	1.00

```
Sample Size
[1] 32
Probability values (Entries above the diagonal are adjusted for multiple tests.)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	0.00	0	0.00	0.00	0.00	0.00	0.10	0.00	0.02	0.03	0.00
cyl	0.00	0	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.02	0.01
disp	0.00	0	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.01	0.03
hp	0.00	0	0.00	0.00	0.04	0.00	0.00	0.00	0.46	0.64	0.00
drat	0.00	0	0.00	0.00	0.00	0.00	1.00	0.12	0.00	0.00	1.00
wt	0.00	0	0.00	0.00	0.00	0.00	1.00	0.01	0.00	0.00	0.05
qsec	0.01	0	0.01	0.00	0.62	0.21	0.00	0.00	1.00	1.00	0.00
vs	0.00	0	0.00	0.00	0.01	0.00	0.00	0.00	1.00	1.00	0.00
am	0.00	0	0.00	0.04	0.00	0.00	0.26	0.36	0.00	0.00	1.00
gear	0.00	0	0.00	0.06	0.00	0.00	0.42	0.12	0.00	0.00	1.00
carb	0.00	0	0.00	0.00	0.49	0.00	0.00	0.00	0.73	0.53	0.00

corr.test(mtcars, method="kendall")

```
> corr.test(mtcars, method="kendall")
Call:corr.test(x = mtcars, method = "kendall")
Correlation matrix
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.80	-0.77	-0.74	0.46	-0.73	0.32	0.59	0.47	0.43	-0.50
cyl	-0.80	1.00	0.81	0.79	-0.55	0.73	-0.45	-0.77	-0.49	-0.51	0.47
disp	-0.77	0.81	1.00	0.67	-0.50	0.74	-0.30	-0.60	-0.52	-0.48	0.41
hp	-0.74	0.79	0.67	1.00	-0.38	0.61	-0.47	-0.63	-0.30	-0.28	0.60
drat	0.46	-0.55	-0.50	-0.38	1.00	-0.55	0.03	0.38	0.58	0.58	-0.10
wt	-0.73	0.73	0.74	0.61	-0.55	1.00	-0.14	-0.49	-0.61	-0.54	0.37
qsec	0.32	-0.45	-0.30	-0.47	0.03	-0.14	1.00	0.66	-0.17	-0.09	-0.51
vs	0.59	-0.77	-0.60	-0.63	0.38	-0.49	0.66	1.00	0.17	0.27	-0.58
am	0.47	-0.49	-0.52	-0.30	0.58	-0.61	-0.17	0.17	1.00	0.77	-0.06
gear	0.43	-0.51	-0.48	-0.28	0.58	-0.54	-0.09	0.27	0.77	1.00	0.10
carb	-0.50	0.47	0.41	0.60	-0.10	0.37	-0.51	-0.58	-0.06	0.10	1.00

```
Sample size
[1] 32
Probability values (Entries above the diagonal are adjusted for multiple tests.)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	0.00	0.00	0.00	0.00	0.15	0.00	1.00	0.01	0.15	0.24	0.09
cyl	0.00	0.00	0.00	0.00	0.04	0.00	0.19	0.00	0.10	0.08	0.15
disp	0.00	0.00	0.00	0.00	0.10	0.00	1.00	0.01	0.07	0.14	0.32
hp	0.00	0.00	0.00	0.00	0.49	0.01	0.14	0.00	1.00	1.00	0.01
drat	0.01	0.00	0.00	0.03	0.00	0.04	1.00	0.52	0.02	0.02	1.00
wt	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.11	0.01	0.04	0.52
qsec	0.08	0.01	0.09	0.01	0.86	0.44	0.00	0.00	1.00	1.00	0.09
vs	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	1.00	1.00	0.02
am	0.01	0.00	0.00	0.09	0.00	0.00	0.36	0.36	0.00	0.00	1.00
gear	0.01	0.00	0.01	0.12	0.00	0.00	0.62	0.14	0.00	0.00	1.00
carb	0.00	0.01	0.02	0.00	0.60	0.04	0.00	0.00	0.75	0.59	0.00

Тесты Стьюдента

Критерий Стьюдента (t-тест)

- **Критерий Стьюдента (t-тест)** - это статистический метод, который позволяет сравнивать средние значения двух выборок и на основе результатов теста делать заключение о том, различаются ли они друг от друга статистически или нет.
- `t.test(y ~ x, data)`
 - y – это числовая переменная,
 - x – дихотомическая
- `t.test(y1, y2)`
 - $y1$ и $y2$ – это числовые векторы (анализируемые значения для каждой из групп).
- Необязательный аргумент `data` назначает матрицу или таблицу данных, в которой содержатся данные.

UScrime

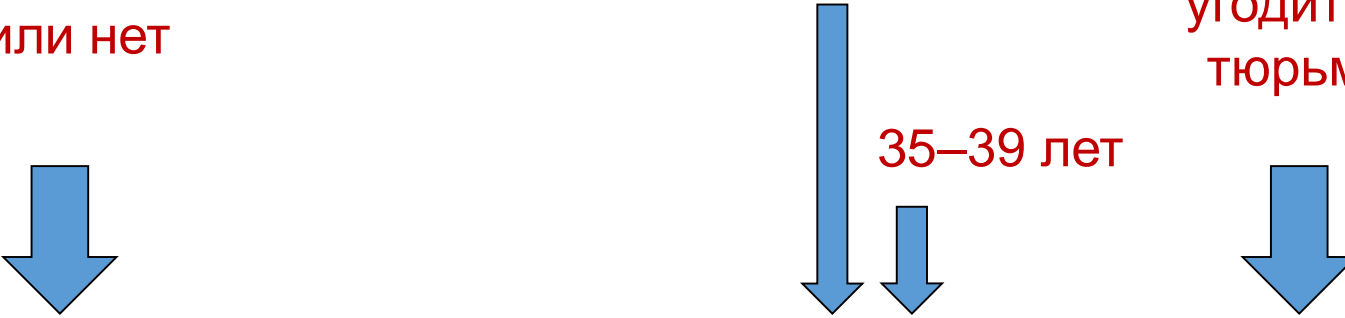
```
install.packages("MASS")  
library(MASS)  
UScrime
```

уровень безработицы
для городских жителей
мужского пола в
возрасте от 14 до 24 лет

южный штат
или нет

вероятность
угодить в
тюрьму

35–39 лет



	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time	y
1	151	1	91	58	56	510	950	33	301	108	41	394	261	0.084602	26.2011	791
2	143	0	113	103	95	583	1012	13	102	96	36	557	194	0.029599	25.2999	1635
3	142	1	89	45	44	533	969	18	219	94	33	318	250	0.083401	24.3006	578
4	136	0	121	149	141	577	994	157	80	102	39	673	167	0.015801	29.9012	1969
5	141	0	121	109	101	591	985	18	30	91	20	578	174	0.041399	21.2998	1234
6	121	0	110	118	115	547	964	25	44	84	29	689	126	0.034201	20.9995	682
7	127	1	111	82	79	519	982	4	139	97	38	620	168	0.042100	20.6993	963
8	131	1	109	115	109	542	969	50	179	79	35	472	206	0.040099	24.5988	1555
9	157	1	90	65	62	553	955	39	286	81	28	421	239	0.071697	29.4001	856

Тест Стьюдента для независимых выборок

Сравнение вероятности попасть в тюрьму в южных штатах ($So=1$) и остальных

```
> t.test(Prob ~ So, data=UScrime)
```

```
Welch Two Sample t-test
```

```
data: Prob by So
```

```
t = -3.8954, df = 24.925, p-value = 0.0006506
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.03852569 -0.01187439
```

```
sample estimates:
```

```
mean in group 0 mean in group 1
```

```
0.03851265
```

```
0.06371269
```

Тест Стьюдента для зависимых выборок

Сравнить уровень безработицы у юношей (14–24 года) и у мужчин (35–39 лет)

```
t.test(y1, y2, paired=TRUE)
```

`paired = TRUE` применение
парного критерия Стьюдента

```
> t.test(UScrime$U1, UScrime$U2, paired=TRUE)
```

Paired t-test

```
data: UScrime$U1 and UScrime$U2
```

```
t = 32.407, df = 46, p-value < 2.2e-16
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
57.67003 65.30870
```

```
sample estimates:
```

```
mean of the differences
```

```
61.48936
```

Разность средних (61.5) достаточно велика, отклонение гипотезы о равенстве уровня безработицы для юношей и мужчин (у юношей она выше).

Визуализация

Кореллограммы

Кореллограммы

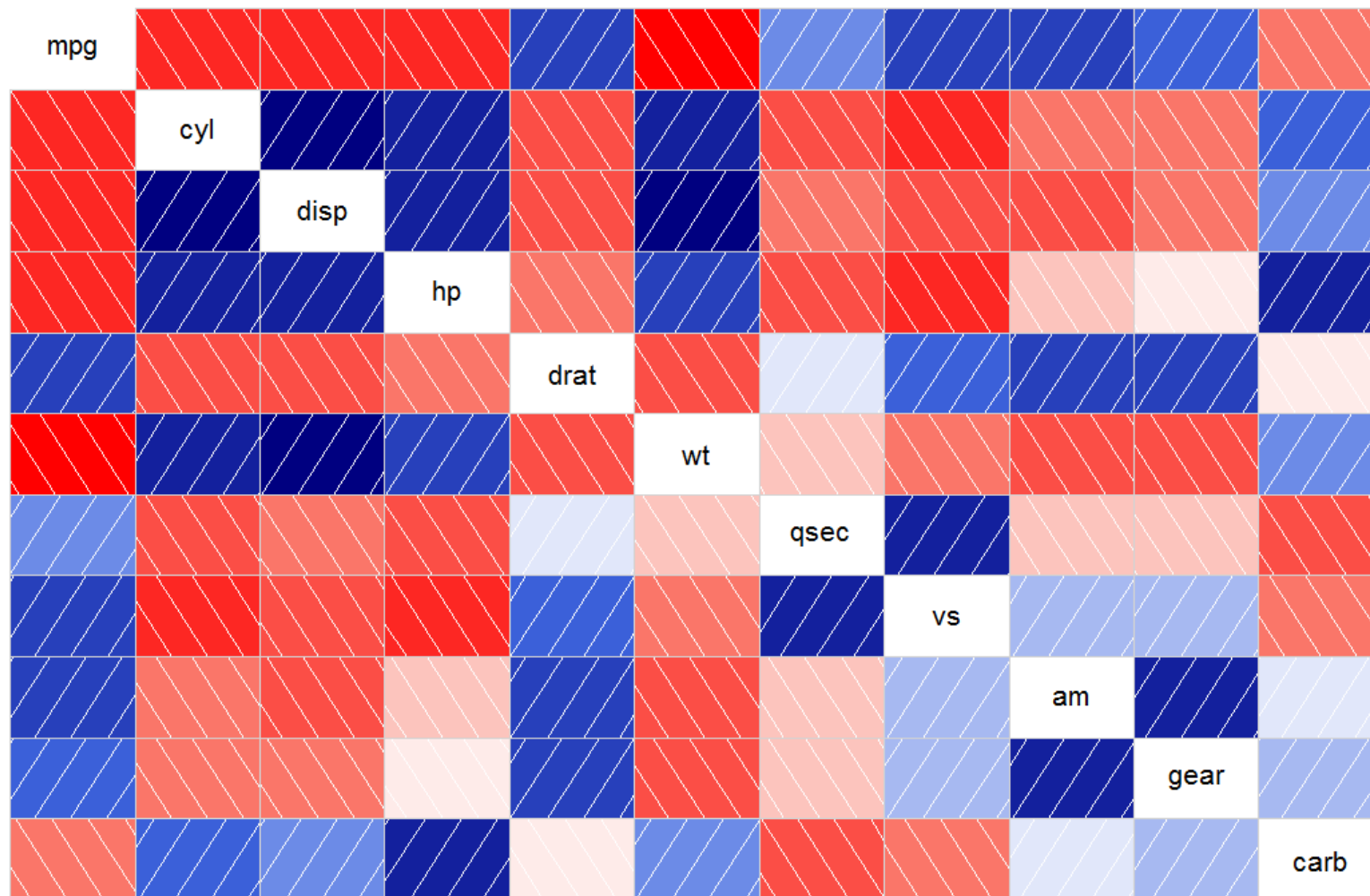
- Матрицы корреляции – это один из основных элементов многомерной статистики. Какие переменные из рассматриваемых сильно коррелируют друг с другом, а какие – нет? Существуют ли кластеры переменных, которые связаны между собой определенным способом? С увеличением числа переменных ответить на такие вопросы становится все сложнее.
- Кореллограммы – это способ для визуализации корреляционных матриц.
- Пример – `corr.test(mtcars, method="pearson")`

```
Correlation matrix
      mpg    cyl  disp    hp  drat    wt   qsec    vs    am  gear   carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

Аннотация к corrgram(). пакет corrgram

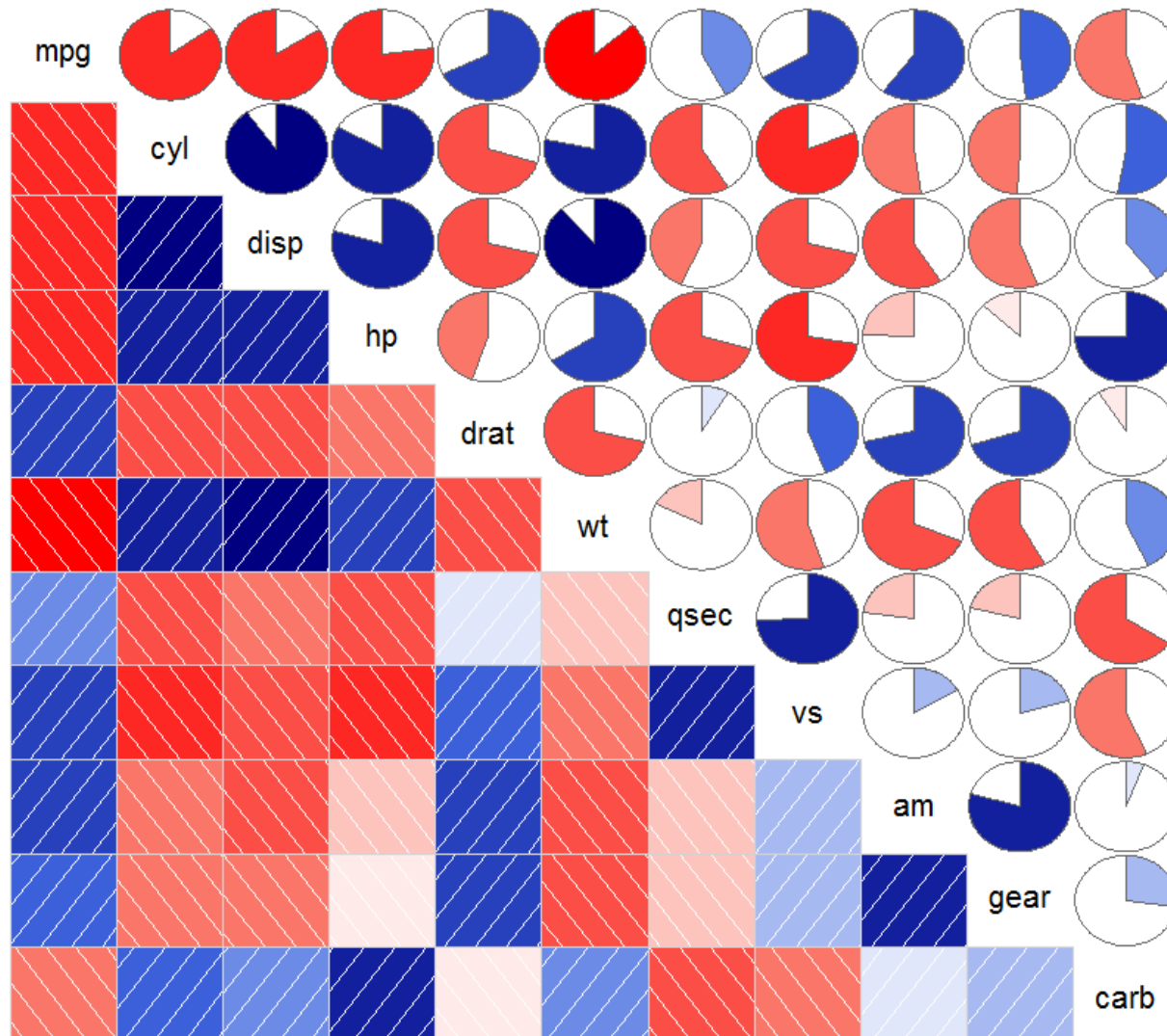
- По умолчанию голубой цвет и штриховка из левого нижнего угла к правому верхнему соответствуют положительной корреляции между двумя переменными, на пересечении которых находится данная ячейка.
- Напротив, красный цвет и штриховка из верхнего левого угла к правому нижнему соответствуют отрицательной корреляции. Чем темнее и насыщеннее цвет, тем сильнее корреляция.
- Слабые, близкие к нулю корреляции будут представлены “выцветшими” ячейками. На представленной диаграмме порядок строк и столбцов был автоматически изменен по результатам анализа главных компонент так, чтобы переменные со сходной корреляционной структурой формировали кластеры.

corrgram(). пакет corrgram



corrgram(). пакет corrgram

corrgram(mtcars, lower.panel=panel.shade, upper.panel=panel.pie)



Аннотация к corrgram()

- На верхнем треугольнике диаграммы та же информация представлена в виде круговых диаграмм.
- Цвета имеют такое же значение, а сила корреляции выражена в размере закрашенного сегмента круговой диаграммы.
- Сегменты, соответствующие положительным корреляциям, начинаются от положения «12 часов» и заполняют круг по часовой стрелке.
- Сегменты, соответствующие отрицательным корреляциям, заполняют круг против часовой стрелки.

Параметры

- `corrgram(x, order=, panel=, text.panel=, diag.panel=)`,
- `x` – это таблица данных с одним наблюдением на строку.
- Если `order=TRUE`, то порядок переменных изменяется согласно результатам анализа главных компонент корреляционной матрицы
- Параметр `panel` определяет вид диаграммы (кроме главной диагонали – ее свойства задаются отдельно). Вместо него можно использовать параметры `lower.panel` и `upper.panel`, чтобы отдельно определять вид нижней и верхней (по отношению к главной диагонали) половин диаграммы.
- Параметры `text.panel` и `diag.panel` относятся к главной диагонали.

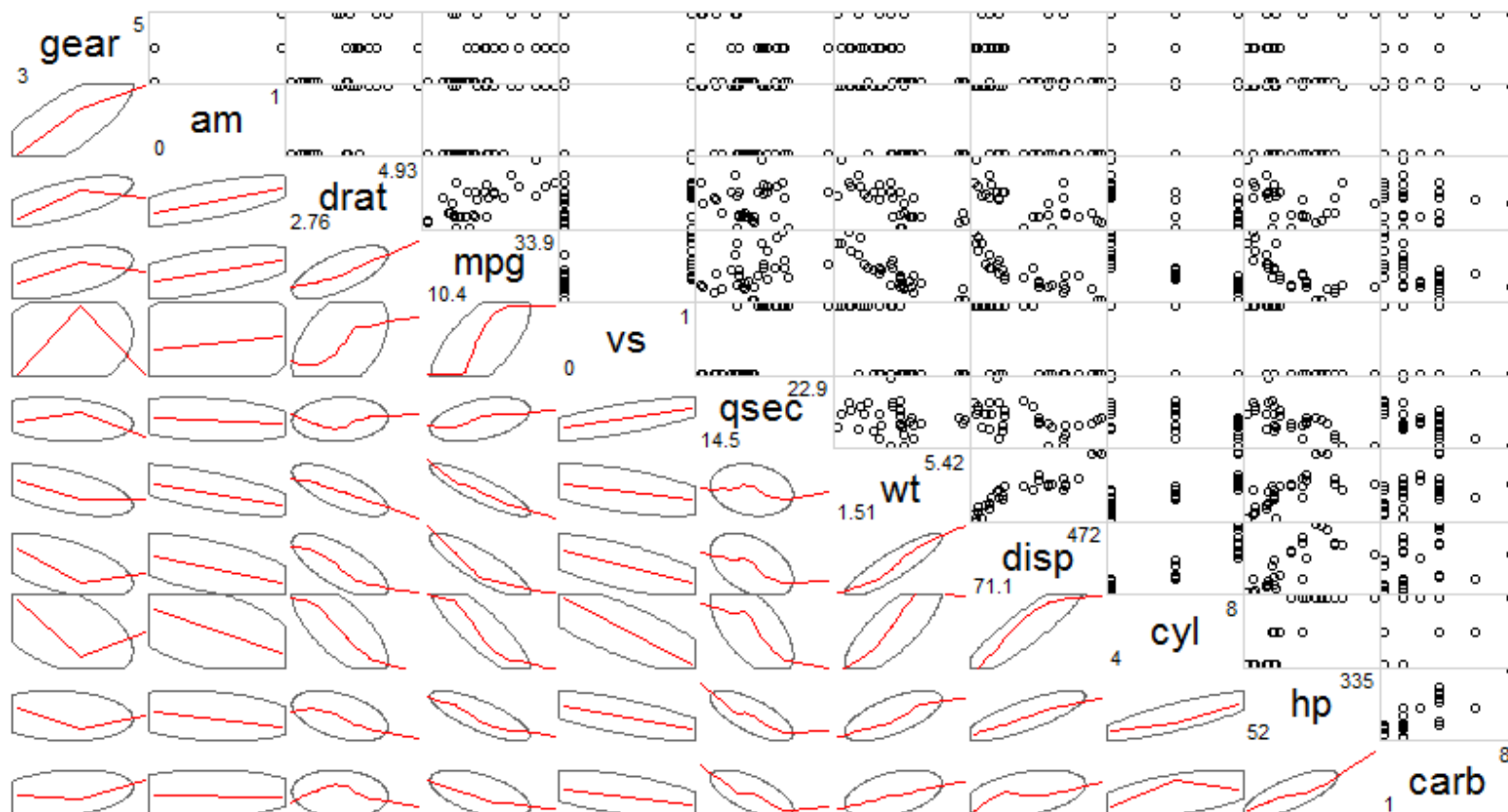
Допустимые значения параметра panel

Положение	Значение параметра	Описание
Не на главной диагонали lower.panel= upper.panel=	panel.pie	Закрашенный сегмент круговой диаграммы соответствует силе корреляции
	panel.shade	Интенсивность цвета соответствует силе корреляции
	panel.ellipse	Изображаются доверительный эллипс и сглаженная линия
	panel.pts	Изображается диаграмма рассеяния
diag.panel=	panel.minmax	Приводятся минимальное и максимальное значения переменной
text.panel=	panel.txt	Отображается название переменной

Параметры

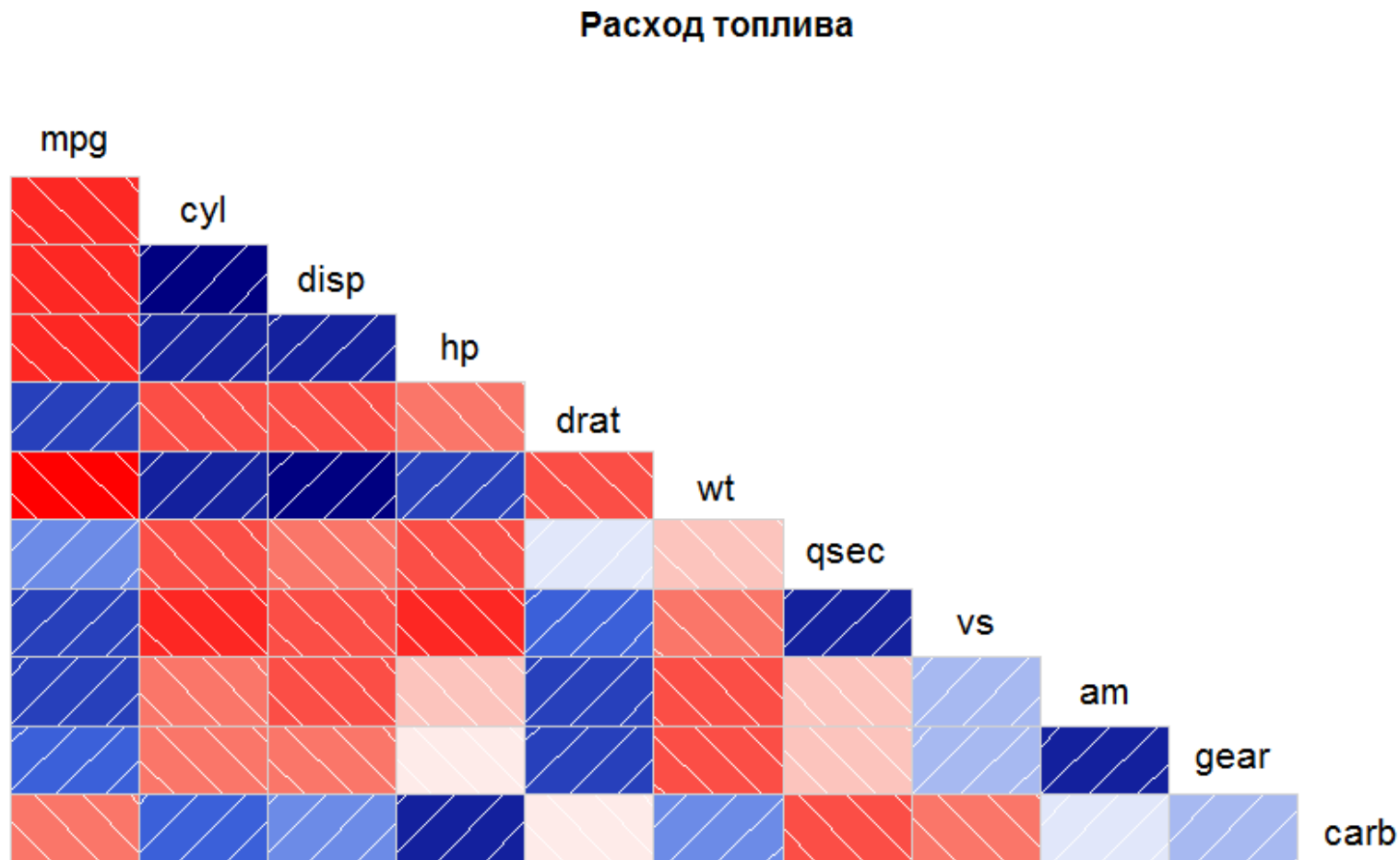
```
corrgram(mtcars, order=TRUE,  
lower.panel=panel.ellipse,upper.panel=panel.pts,  
text.panel=panel.txt,diag.panel=panel.minmax,main="Кореллограмма")
```

Кореллограмма



Параметры

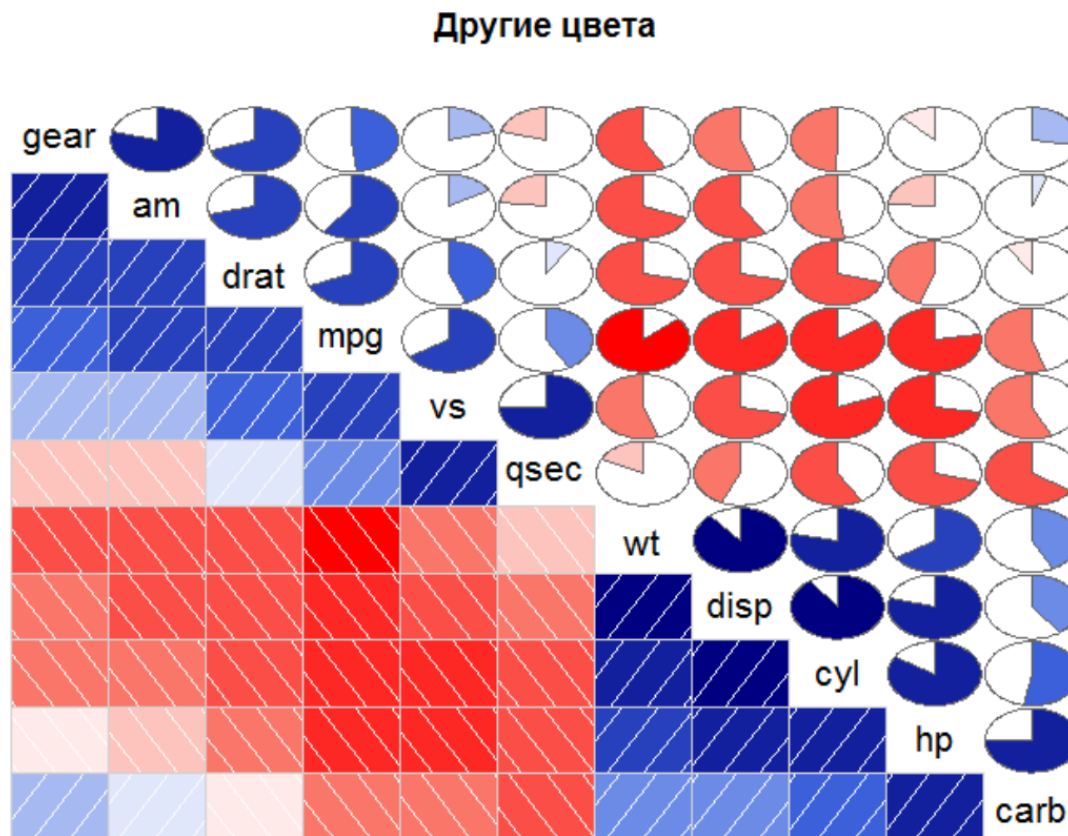
```
corrgram(mtcars, lower.panel=panel.shade, upper.panel=NULL,  
text.panel=panel.txt, main="Расход топлива")
```



Управление цветом. Определяются 4 цвета

```
col.corrgram <- function(ncol){colorRampPalette(c("darkgoldenrod4",  
"burlywood1", "darkkhaki", "darkgreen"))(ncol)}
```

```
corrgram(mtcars, order=TRUE, lower.panel=panel.shade,  
upper.panel=panel.pie, text.panel=panel.txt, main="Другие цвета")
```

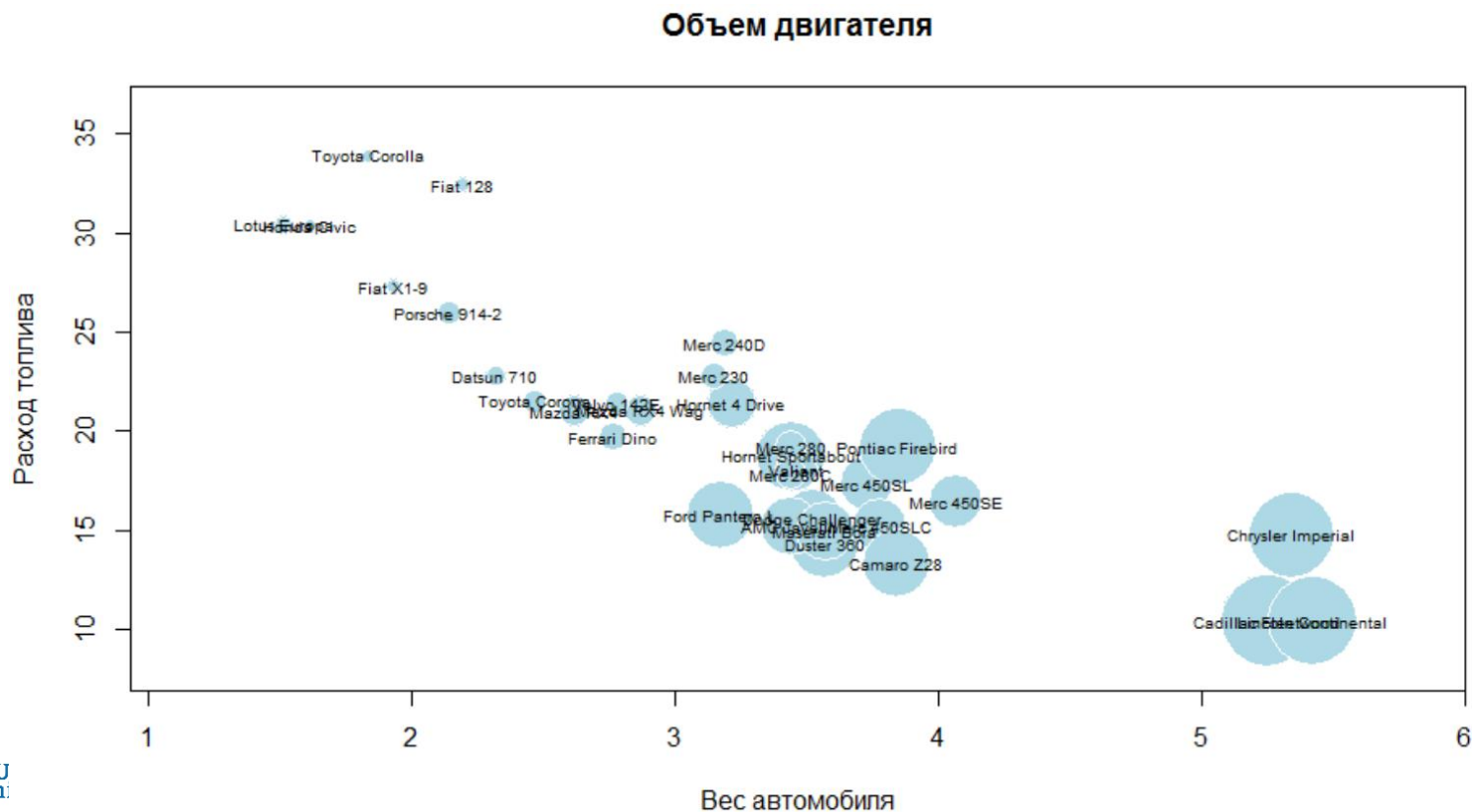


Пузырьковые диаграммы

symbols()

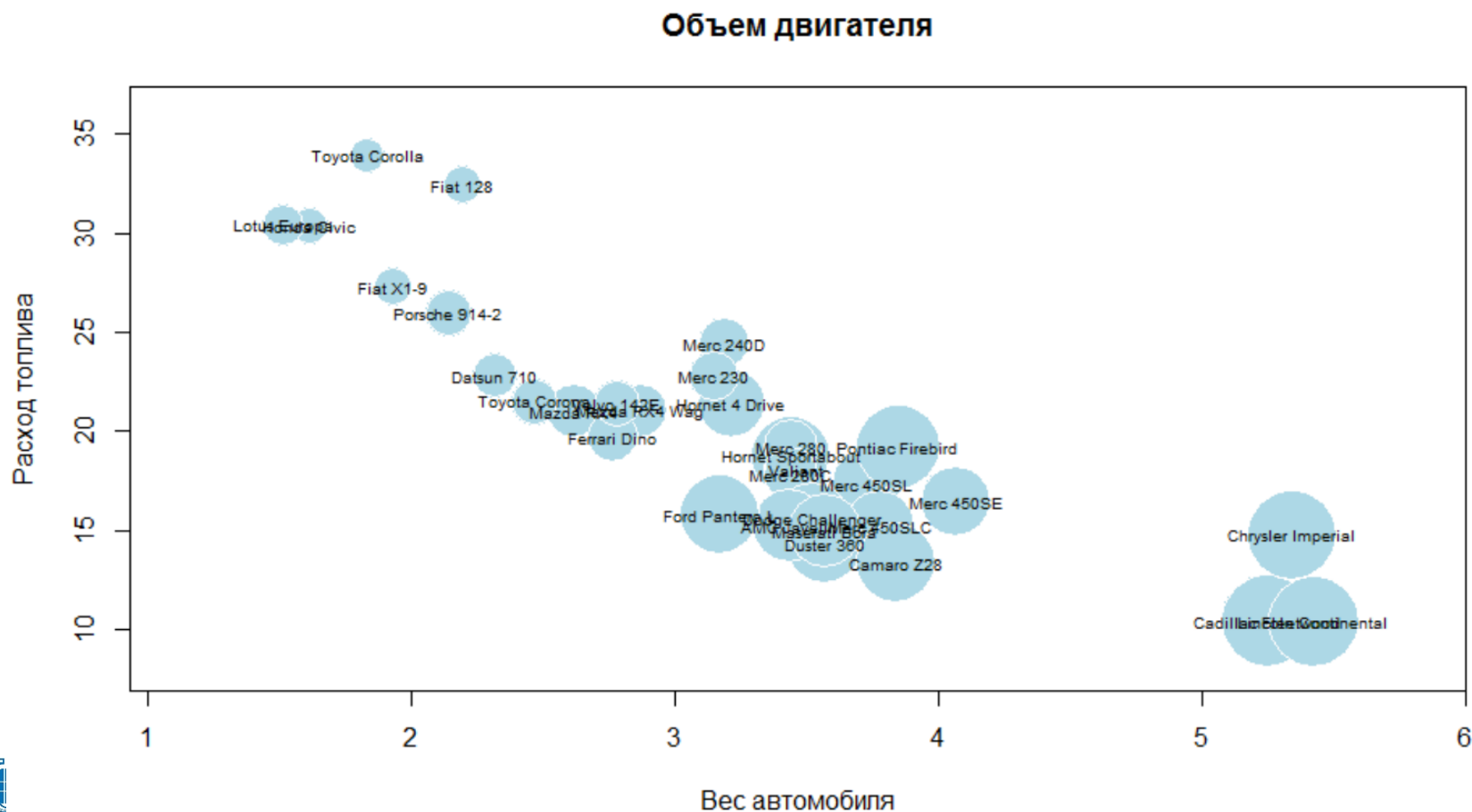
```
symbols(mtcars$wt, mtcars$mpg, circle=mtcars$disp,  
inches=0.30,fg="white",bg="lightblue",main="Объем двигателя",  
ylab="Расход топлива",xlab="Вес автомобиля")
```

```
text(mtcars$wt, mtcars$mpg, rownames(mtcars), cex=0.6)
```



symbols()

```
symbols(mtcars$wt, mtcars$mpg, circle=sqrt(mtcars$disp/pi),  
inches=0.30,fg="white",bg="lightblue",main="Объем  
двигателя",ylab="Расход топлива",xlab="Вес автомобиля")  
text(mtcars$wt, mtcars$mpg, rownames(mtcars), cex=0.6)
```



Мозаичные диаграммы

Мозаичные диаграммы

- Применяются для набора категориальных переменных
- Одна категориальная переменная - столбчатая или круговая диаграмма
- Две - трехмерная столбчатая диаграмма
- Более - Мозаичная диаграмма
- В мозаичных диаграммах частоты из многомерной таблицы сопряженности представлены в виде вложенных прямоугольников, размер которых пропорционален частотам.
- базовый пакет - `mosaicplot()`
- пакет `vcd` - `mosaic()`

Titanic

Содержит данные по выжившим по разным категориям (палуба, пол, возраст)

```
> Titanic
```

```
, , Age = Child, Survived = No
```

	Sex	
Class	Male	Female
1st	0	0
2nd	0	0
3rd	35	17
Crew	0	0

```
, , Age = Adult, Survived = No
```

	Sex	
Class	Male	Female
1st	118	4
2nd	154	13
3rd	387	89
Crew	670	3

```
, , Age = Child, Survived = Yes
```


Titanic

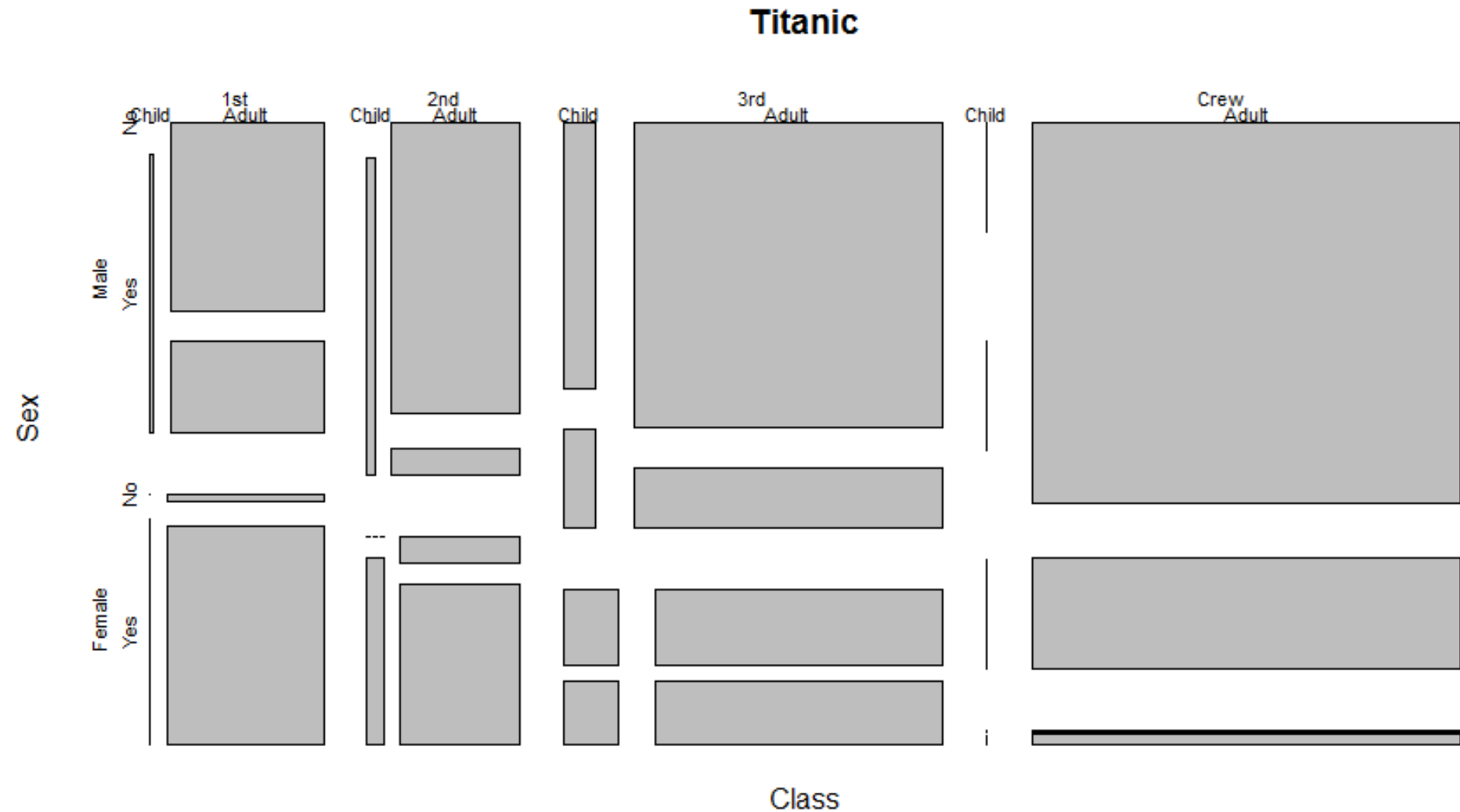
Подсчет числа сочетаний разных признаков

```
> ftable(Titanic)
```

			Survived	No	Yes
Class	Sex	Age			
1st	Male	Child		0	5
		Adult		118	57
	Female	Child		0	1
		Adult		4	140
2nd	Male	Child		0	11
		Adult		154	14
	Female	Child		0	13
		Adult		13	80
3rd	Male	Child		35	13
		Adult		387	75
	Female	Child		17	14
		Adult		89	76
Crew	Male	Child		0	0
		Adult		670	192
	Female	Child		0	0
		Adult		3	20

mosaicplot()

mosaicplot(Titanic)

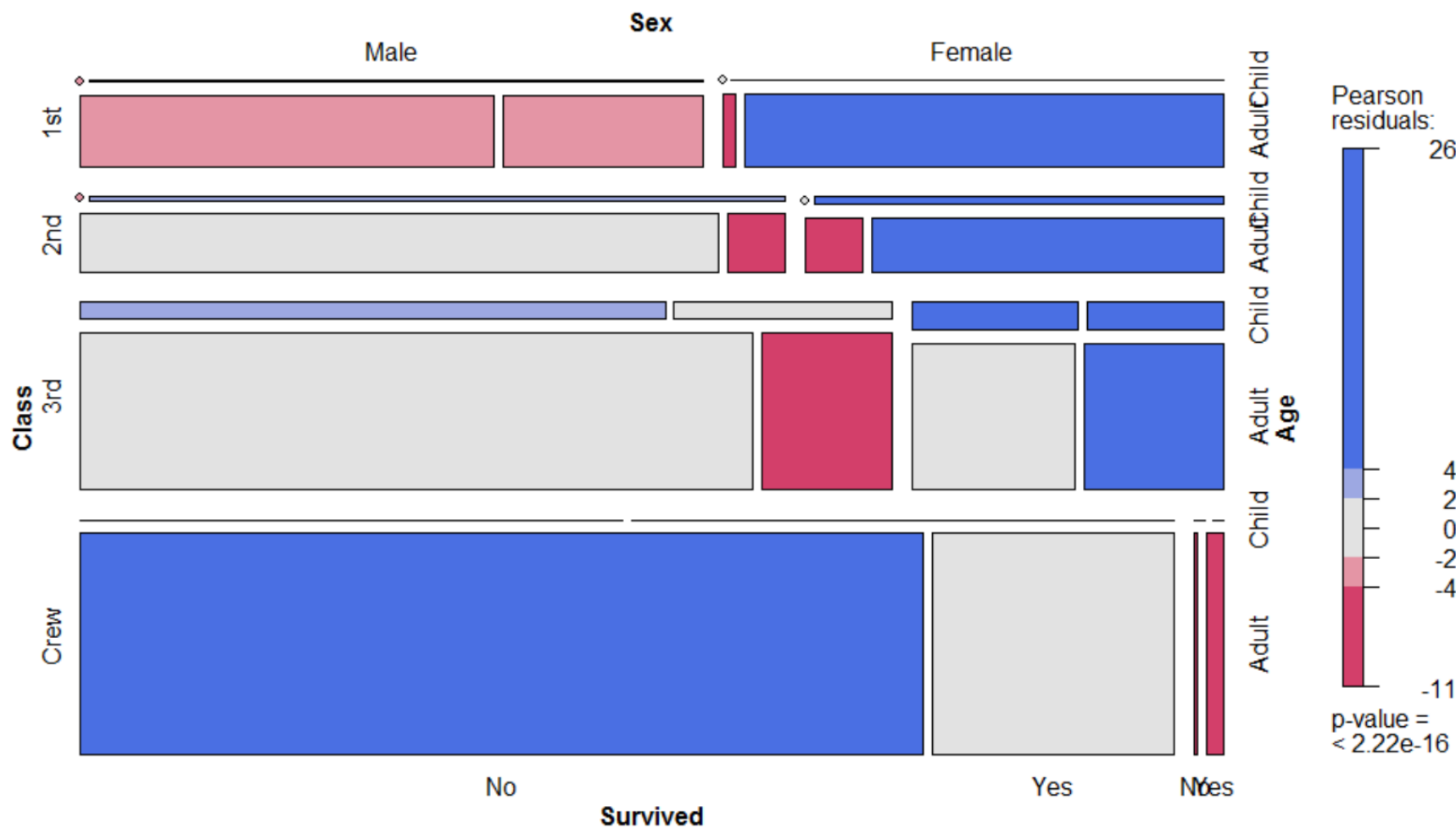


mosaic()

```
install.packages("vcd")
```

```
library(vcd)
```

```
mosaic(Titanic, shade=TRUE, legend=TRUE)
```



Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57