#Лабораторный практикум ВКИАД с R

Преподаватель Дорошко Ольга Валерьевна

Студент Черепенников Роман Михайлович, 2 курс, 8 группа, ПМ

Вариант № 13

Часть 1. Управление данными в R

Ввести данные из текстового файла согласно варианту (функция “read.table”). Сохранить в переменную qc (таблица данных).

```
qc <- read.table("QC13.txt")
```

Вывести размерность таблицы данных(dim).

```
dim(qc)
```

```
## [1] 100   9
```

Вывести первые 3 записи таблицы (head).

```
head(qc,3)
```

```
##      V1    V2 V3 V4 V5 V6     V7    V8 V9
## 1 50.75  9.13  2  2  1  4 16.644 46.77  4
## 2 54.70  9.11  2  3  1  4 18.210 47.82  3
## 3 57.66 11.53  5  1  1  4 19.536 46.44  3
```

Изменить имена переменных в таблице (names) согласно: “б”-“kit”, “п”-“nbug”, “№изг”-“maker”, “№пост”-“vendor”, “№вд”-“bug”, “н” - последовательно проименовать как “c1”, “c2”, “c3”, “c4” (порядок переменных смотрите в файле Lab.pdf согласно Вашему варианту).
Вывести часть таблицы функцией head.

```
colnames(qc) = c("c1","c2","kit","nbug","maker","vendor","c3","c4","bug")
head(qc)
```

```
##      c1    c2 kit nbug maker vendor     c3    c4 bug
## 1 50.75  9.13   2    2     1      4 16.644 46.77   4
## 2 54.70  9.11   2    3     1      4 18.210 47.82   3
## 3 57.66 11.53   5    1     1      4 19.536 46.44   3
## 4 56.43  9.45   1    1     1      3 18.783 45.63   5
## 5 52.69  8.40   6    2     2      3 17.760 49.14   3
## 6 55.26 10.73   2    5     1      2 18.675 48.84   6
```

Вывести общую статистику по переменной vendor (summary). Обращаться к переменной через “$”.

```
summary(qc$vendor)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    3.00    2.87    4.00    4.00
```

Преобразовать переменную vendor в переменную-фактор (с помощью as.factor, для доступа к переменной обратиться к столбцу с помощью оператора [] или оператора [[]]).
Вывести общую статистику для преобразованной переменной vendor.

```
summary(as.factor(qc[,"vendor"]))
```

```
##  1  2  3  4
## 14 23 25 38
```

Подсчитать количество изделий, у которых число некритических дефектов больше 3 функция sum и оператор условия).

```
sum(qc$nbug > 3)
```

```
## [1] 14
```

Вывести наблюдение(-я), для которого количество некритических дефектов равно максимальному (оператор [] к переменной qc, функция max).

```
qc[which.max(qc[,"nbug"]),]
```

```
##       c1   c2 kit nbug maker vendor     c3     c4 bug
## 93 53.14 9.12   5    7     2      4 17.877 56.025   6
```

Часть 2. Описательная статистика и графический анализ

Присоединить таблицу данных qc к списку текущих переменных (attach). Далее обращаться к именам переменных напрямую.
Загрузить дополнительную библиотеку (для вычисления коэффициентов асимметрии и эксцесса).

```
attach(qc)
library("e1071")
```

В данном пункте статистики и графики выводятся только для переменной c1.

Вычислить среднее (mean) и медиану (median).

```
mean(c1)
```

```
## [1] 54.7083
```

```
median(c1)
```

```
## [1] 54.92
```

Вычислить выборочные квартили Q1, Q2, Q3 (quantile) (в одну строчку). Вычислить интерквартильный размах с помощью встроенной в R функции.

```
quantile(qc$c1,c(0.25,0.5,0.75))
```

```
##     25%     50%     75%
## 53.5075 54.9200 56.2775
```

```
IQR(c1)
```

```
## [1] 2.77
```

Вывести комплексную статистику по переменной c1.

```
summary(c1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.86   53.51   54.92   54.71   56.28   57.96
```

Вычислить дисперсию (var) для переменных c1 и 2 * c1 + 100 в указанном порядке. В выводе записать программно вычисленное отношение дисперсий данных переменных.

```
var(c1)

## [1] 3.414362

var(c1*2+100)

## [1] 13.65745

var(c1*2+100)/var(c1)

## [1] 4
```

Вывод: вторая дисперсия в четыре раза больше

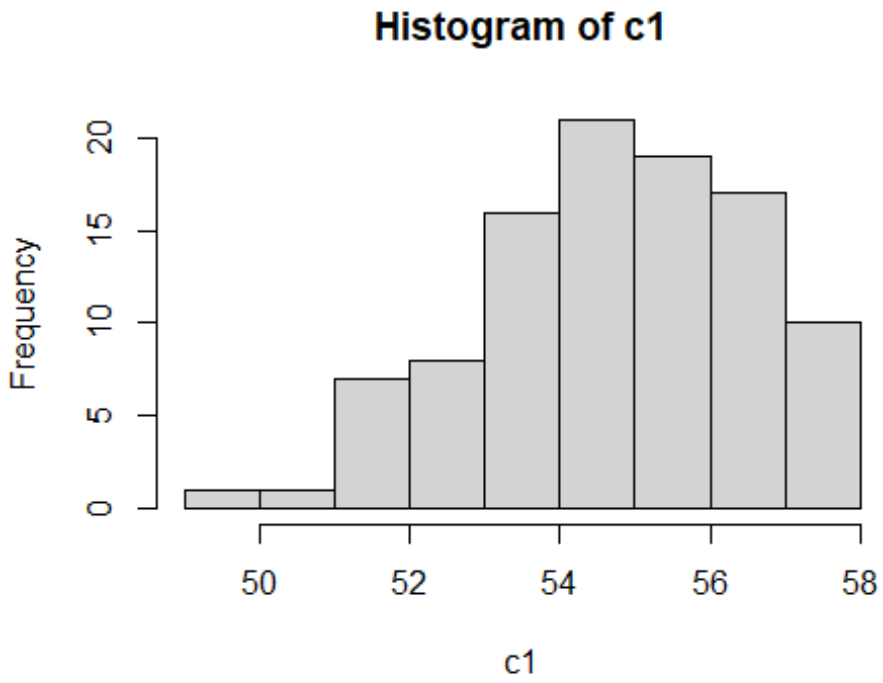
Вычислить дисперсию, используя векторное выражение (то есть в одну строку, без циклов).

```
sum((c1-mean(c1))^2)/(length(c1)-1)

## [1] 3.414362
```

Построить гистограмму (hist).

```
hist(c1)
```



Вывести асимметрию (skewness) и эксцесс (kurtosis) в указанном порядке. Сделать вывод.

```
skewness(c1)
```

```
## [1] -0.4037598
```

```
kurtosis(c1)
```

```
## [1] -0.4987426
```

Вывод: распределение c1 скошено влево и вершина смещена вниз

Проверить нормальность остатков с помощью статистического теста Лиллиефорса (lillie.test). Сделать вывод.

```
library(nortest)
lillie.test(c1)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  c1
## D = 0.060991, p-value = 0.478
```

Вывод: гипотеза о нормальном распределении не отклоняется на уровне значимости 0.05

Часть 3. Анализ статистических зависимостей

Перекодировать переменную kit так, чтобы она принимала только два значения: 1 и 2 (1 при kit < 3, 2 - в отстальных случаях).

```
qc$kit <- ifelse(qc$kit>3,2,1)
print(qc)
```

```
##        c1    c2 kit nbug maker vendor     c3     c4 bug
## 1   50.75  9.13   1    2     1      4 16.644 46.770   4
## 2   54.70  9.11   1    3     1      4 18.210 47.820   3
## 3   57.66 11.53   2    1     1      4 19.536 46.440   3
## 4   56.43  9.45   1    1     1      3 18.783 45.630   5
## 5   52.69  8.40   2    2     2      3 17.760 49.140   3
## 6   55.26 10.73   1    5     1      2 18.675 48.840   6
## 7   54.99  9.80   1    2     1      4 18.594 47.550   3
## 8   54.05 10.10   1    4     2      4 18.393 49.485   3
## 9   53.79 10.40   1    1     1      3 18.393 47.640   2
## 10  51.01  9.99   1    3     2      3 17.169 50.445   3
## 11  56.88 10.44   2    1     2      4 18.849 51.930   3
## 12  52.17  8.50   1    3     2      4 17.154 52.185   4
## 13  51.14  8.70   1    3     2      2 16.593 50.115   6
## 14  51.85 11.16   2    2     2      4 17.097 50.265   3
## 15  56.99 11.16   1    1     1      3 18.948 47.265   3
## 16  57.22 10.45   1    2     1      3 19.257 46.920   4
## 17  54.96 11.06   1    3     2      3 17.970 53.370   5
## 18  54.76  8.83   1    1     1      3 18.312 50.100   2
## 19  54.11 10.62   1    0     1      1 17.898 48.390   5
## 20  54.77 10.11   1    4     1      1 18.318 49.635   2
## 21  54.59  9.23   1    4     1      4 18.063 45.735   5
## 22  57.86 11.36   1    3     1      3 18.294 50.580   4
## 23  52.12 10.06   1    2     2      3 17.220 49.080   3
## 24  55.49 10.13   1    2     2      4 18.495 54.030   4
## 25  52.50  9.98   1    3     1      2 17.565 49.560   2
## 26  54.41  9.94   1    1     1      4 18.393 48.315   3
## 27  56.07 10.65   1    4     2      3 18.504 53.415   3
## 28  55.21 10.79   1    0     2      2 18.504 53.340   4
## 29  56.27 10.59   1    1     1      3 18.717 47.820   3
## 30  52.78  7.78   1    0     1      2 17.178 49.725   4
## 31  52.10  9.34   1    3     1      2 17.463 49.050   5
## 32  56.84 10.30   1    3     1      4 18.969 45.960   3
## 33  52.63  9.70   1    1     1      2 17.406 47.835   2
## 34  56.45  9.13   2    4     1      4 18.762 46.560   2
## 35  53.50  9.47   1    1     1      4 18.207 50.205   4
## 36  54.46  9.50   1    3     1      4 17.475 49.680   2
## 37  57.41 11.09   1    4     2      4 19.722 54.315   4
## 38  54.06  9.79   1    2     2      3 18.564 52.665   3
## 39  54.65 10.13   1    2     2      2 18.528 54.525   3
## 40  53.25  9.87   1    1     1      4 17.670 45.555   2
## 41  53.58  9.96   1    3     2      2 17.457 47.850   6
## 42  55.35  9.61   2    1     2      1 18.150 48.780   3
## 43  55.11  9.60   1    1     1      4 18.570 49.260   3
## 44  55.52 10.94   1    1     1      4 18.429 44.925   3
```

```
## 45  53.23 10.12   1    2     2      4 18.378 48.075   7
## 46  57.20 10.04   1    1     2      4 19.236 48.450   7
## 47  53.51  9.04   1    0     1      3 17.571 50.415   3
## 48  54.69 10.51   1    2     1      2 17.571 49.725   3
## 49  53.57  9.58   2    1     2      2 18.018 51.405   3
## 50  51.59  8.31   1    1     2      1 16.905 49.725   3
## 51  57.63 10.62   1    2     2      2 19.362 51.420   7
## 52  53.34  9.14   1    4     2      2 17.721 52.455   4
## 53  52.96  9.70   1    2     2      3 18.246 49.410   3
## 54  57.66 10.68   2    2     2      1 19.362 53.400   3
## 55  56.39 10.55   1    4     2      3 18.189 52.695   5
## 56  53.14  9.39   2    2     2      4 17.541 49.260   4
## 57  54.84  9.28   1    6     2      2 18.339 52.650   3
## 58  55.65 10.69   2    2     1      3 18.636 47.760   2
## 59  53.80 10.19   1    3     1      4 17.850 51.075   6
## 60  54.91  9.77   1    2     1      3 18.132 47.325   3
## 61  55.82  9.36   2    1     1      4 17.886 47.910   2
## 62  54.93 10.54   1    0     2      1 18.783 53.685   3
## 63  51.02  9.79   1    0     2      3 16.746 50.340   5
## 64  56.57 11.15   2    3     1      4 18.936 48.360   2
## 65  53.60  9.14   2    0     1      2 17.745 48.780   6
## 66  56.00  9.82   2    2     1      3 18.654 46.140   3
## 67  56.33 11.01   2    0     1      4 19.116 46.755   4
## 68  51.20 10.20   2    2     1      4 17.460 46.395   5
## 69  57.04 10.22   1    2     1      2 18.960 52.665   3
## 70  56.48 10.47   1    1     1      1 19.017 46.140   3
## 71  53.14  9.77   1    0     2      4 17.856 52.845   3
## 72  54.97 10.73   2    3     2      3 19.125 51.225   5
## 73  54.09  9.49   2    3     2      4 17.868 52.275   7
## 74  51.12  7.29   1    2     1      2 16.719 47.460   2
## 75  57.96 10.76   1    4     2      3 18.360 52.485   7
## 76  55.19 10.68   1    3     2      4 18.606 51.930   3
## 77  55.85 10.94   1    3     1      4 19.038 48.240   2
## 78  55.78  9.51   2    2     2      1 18.477 53.940   4
## 79  56.46 11.44   1    3     1      1 19.101 45.300   2
## 80  56.30 11.89   2    3     2      4 18.771 47.460   3
## 81  56.53 11.41   1    1     2      2 19.419 51.195   4
## 82  55.48  8.72   1    2     1      2 18.162 50.325   2
## 83  55.41 10.72   1    2     2      4 18.096 52.395   5
## 84  49.86  7.43   1    3     2      1 16.539 49.185   4
## 85  53.48  8.96   2    2     1      1 17.766 46.455   2
## 86  56.32  9.23   1    2     1      2 19.119 45.930   2
## 87  56.71 10.22   2    2     1      1 18.288 50.205   6
## 88  54.55 11.45   1    4     2      1 18.264 52.305   4
## 89  55.15 10.14   1    3     1      2 17.907 44.280   2
## 90  55.56 10.13   2    1     1      4 18.048 50.295   2
## 91  55.40  7.74   1    1     2      3 18.234 53.895   3
## 92  54.18 11.73   2    2     1      4 18.552 50.100   6
## 93  53.14  9.12   2    7     2      4 17.877 56.025   6
## 94  53.57 10.93   1    1     1      3 17.931 48.345   5
## 95  55.63  9.92   1    0     2      2 18.567 48.225   3
## 96  53.69  8.71   1    0     2      4 17.931 51.015   4
```

```
## 97  54.17 10.37   2    4     2      2 18.219 50.265   3
## 98  55.84 10.29   1    0     2      1 18.321 49.980   3
## 99  56.67 10.54   1    2     1      3 19.137 46.725   2
## 100 57.14 10.05   2    3     2      4 18.783 50.235   6
```

Вычислить ранговый коэффициент корреляции Спирмена (cor, method = “spearman”) для переменных kit и maker.

```
cor(kit,maker,method = "spearman")

## [1] 0.07044474
```

Вывести таблицу сопряженности (table) по переменным kit и maker. Сделать вывод.

```
table(kit,maker)

##    maker
## kit  1  2
##   0  1  4
##   1 12  7
##   2 16 11
##   3 10 12
##   4  8  6
##   5  4  5
##   6  1  3
```

Вывод: согласно ранговому коэффициенту корреляции и таблице сопряженности согласованность между переменными практически отсутствует

Для переменных c1, c3 вычислить коэффициент корреляции. Сделать вывод.

```
cor.test(c1,c3)

##
##  Pearson's product-moment correlation
##
## data:  c1 and c3
## t = 17.898, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8195065 0.9143318
## sample estimates:
##       cor
## 0.8750684
```

Вывод: гипотеза о равенстве нулю коэффициента корреляции между переменными отклоняется на уровне значимости 0.05.

Вывести корреляционную матрицу по всем числовым характеристикам c1-c4 (cor). Для выбора нужных столбцов таблицы воспользоваться оператором “[]”.

```
cor(qc[c("c1","c2","c3","c4")])

##             c1           c2          c3           c4
## c1 1.000000000  0.539350201  0.87506839  0.003347985
```

```
## c2 0.539350201  1.000000000  0.58341933 -0.009795793
## c3 0.875068389  0.583419330  1.00000000 -0.010034222
## c4 0.003347985 -0.009795793 -0.01003422  1.000000000
```

Представить матрицу корреляции в более удобном виде(symnum, в качестве аргумента данной функции использовать результат функции cor). Сделать вывод.

```
symnum(cor(qc[c("c1","c2","c3","c4")]))
```

```
##    c1 c2 c3 c4
## c1 1
## c2 .  1
## c3 +  .  1
## c4          1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

Вывод: наибольший коэффициент корреляции меэду пременными c1 и c3

Построить модель линейной регрессии для пары переменных, для которых коэффициент корреляции имеет наибольшее значение. Использовать функцию lm, зависимая переменная - с меньшим номером. Полученную модель сохранить в переменную linmod. Вывести общую статистику по модели (summary).

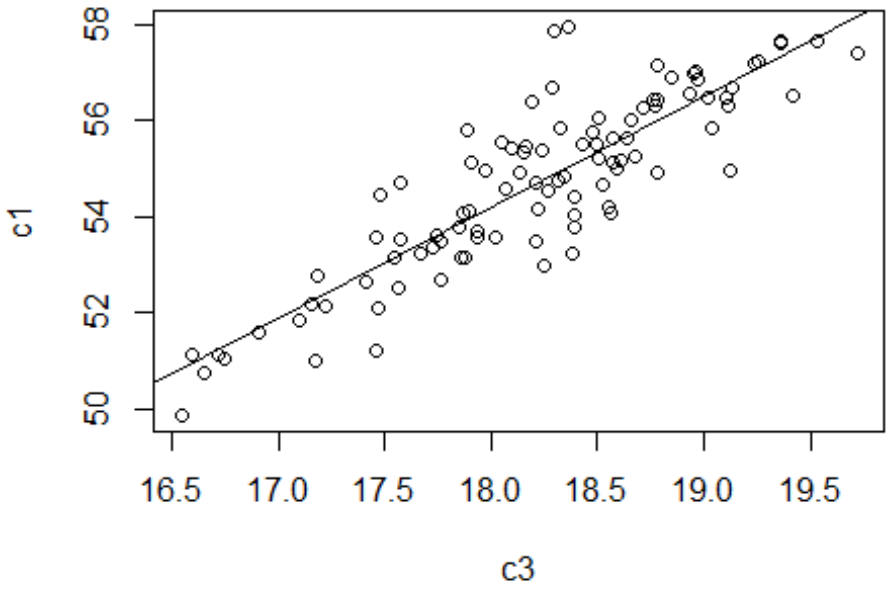
```
linmod=lm(c1~c3)
summary(linmod)
```

```
##
## Call:
## lm(formula = c1 ~ c3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83620 -0.50124 -0.07263  0.42298  2.98774
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.633      2.353    5.37  5.3e-07 ***
## c3             2.309      0.129   17.90  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8989 on 98 degrees of freedom
## Multiple R-squared:  0.7657, Adjusted R-squared:  0.7634
## F-statistic: 320.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

Вывод: Коэффициент при зависимой переменной является статистически значимым (на уровне 0.05). Коэффициент при независимой переменной является статистически значимым (на уровне 0.05). Статистика R-квадрат принимает значение 0.7657.

Построить диаграмму рассеяния (plot) с линией регрессии (abline) для переменных, которые участвуют в оцененной модели.

```
plot(c3,c1,abline(linmod))
```



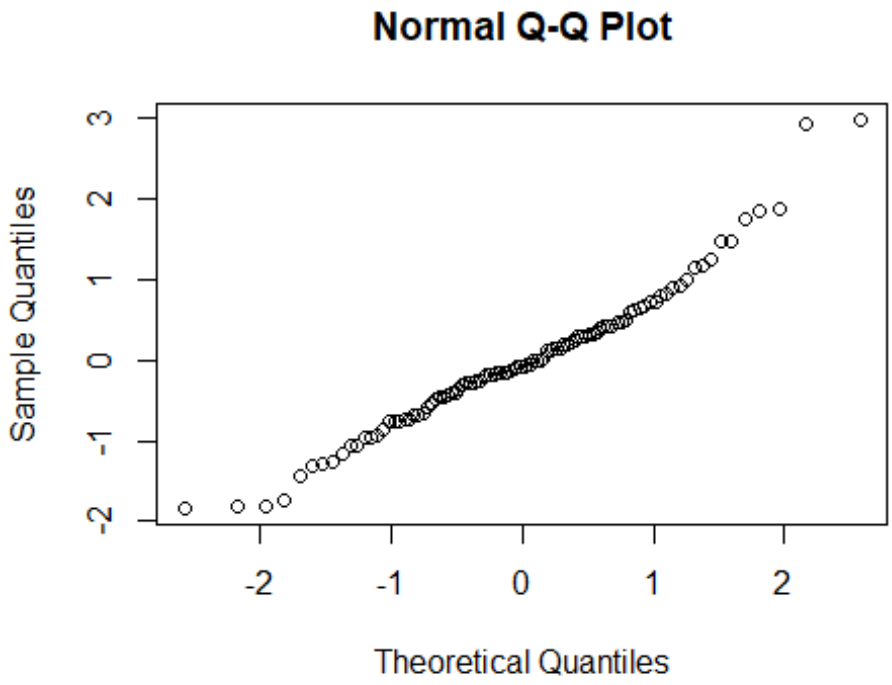
Вывести отдельно коэффициенты модели (coefficients).

```
coefficients(linmod)

## (Intercept)          c3
##   12.632960    2.308915
```

Сохранить значения остатков (residuals) модели linmod в переменную res. Построить график “Квантиль-квантиль” (qqnorm) для остатков. Сделать вывод.

```
res<-residuals(linmod)
qqnorm(res)
```



Вывод: большинство точек располагаются вблизи прямой линии, поэтому распределение остатков близко к нормальному.

Проверить нормальность остатков с помощью статистического теста Колмогорова-Смирнова (ks.test).

```
ks.test(res,pnorm)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  res
## D = 0.099534, p-value = 0.2751
## alternative hypothesis: two-sided
```

Вывод: Критерий принимает гипотезу о нормальном распределении остатков на уровне значимости 0.05

Часть 4. Анализ неоднородных данных

Подсчитать частоты значений для переменной maker (table).

```
table(maker)
```

```
## maker
##  1  2
## 52 48
```

Последовательно применить двухвыборочный t-критерий (t.test) к переменным c1, c2, c3, c4. Ввборки значений каждой переменной разделяются на две подвыборки по

значениям переменной maker (1 или 2).
Программно получить и вывести результаты (статистику по тесту) по той переменной (c1 / c2 / c3 / c4), для которой наблюдается значимое (на уровне 0.05) различие средних в двух подвыборках. Далее анализировать только выбранную переменную.

```
t.test(c1~maker)
```

```
##
##  Welch Two Sample t-test
##
## data:  c1 by maker
## t = 1.1891, df = 93.242, p-value = 0.2374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2957362  1.1786208
## sample estimates:
## mean in group 1 mean in group 2
##        54.92019        54.47875
```

```
t.test(c2~maker)
```

```
##
##  Welch Two Sample t-test
##
## data:  c2 by maker
## t = 0.17399, df = 96.525, p-value = 0.8622
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3379137  0.4028496
## sample estimates:
## mean in group 1 mean in group 2
##        9.997885        9.965417
```

```
t.test(c3~maker)
```

```
##
##  Welch Two Sample t-test
##
## data:  c3 by maker
## t = 0.69076, df = 94.003, p-value = 0.4914
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1828970  0.3780508
## sample estimates:
## mean in group 1 mean in group 2
##        18.26983        18.17225
```

```
t.test(c4~maker)
```

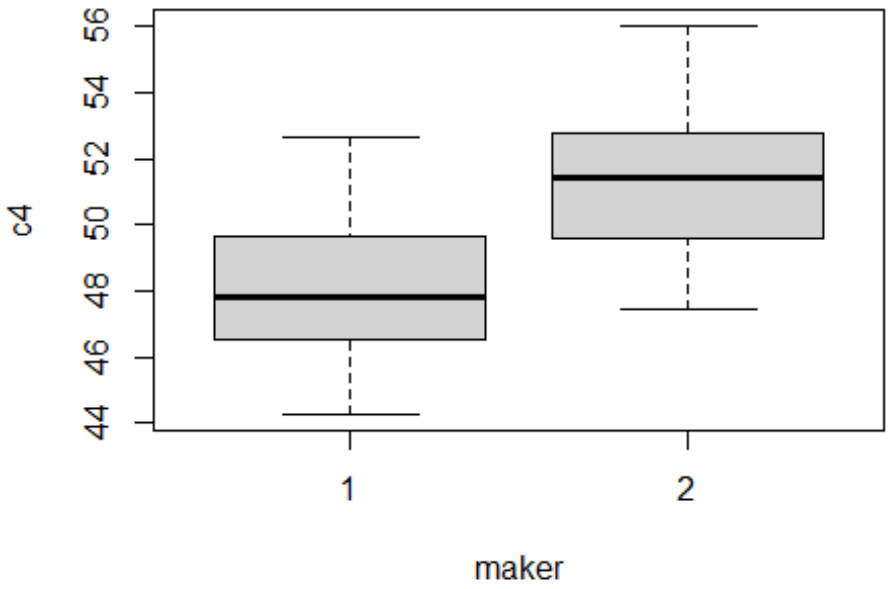
```
##
##  Welch Two Sample t-test
##
## data:  c4 by maker
## t = -8.5391, df = 94.451, p-value = 2.26e-13
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.107890 -2.558024
## sample estimates:
## mean in group 1 mean in group 2
##        48.01673        51.34969
```

Вывод: Гипотеза о равенстве средних в двух подвыборках отклоняется на уровне значимости 0.05 для переменной c4 (средние значения характеристики c4 у двух производителей статистически различимы).

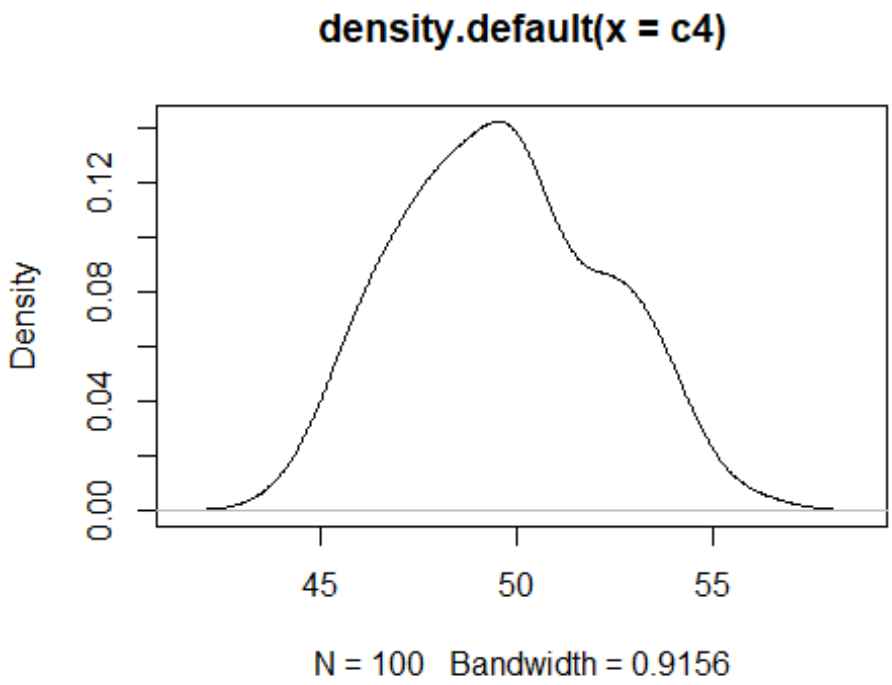
Построить “ящик с усами” для анализируемой переменной в разрезе по номеру производителя (maker).

```
boxplot(c4~maker)
```



Построить график ядерной оценки плотности распределения (density) для анализируемой переменной.

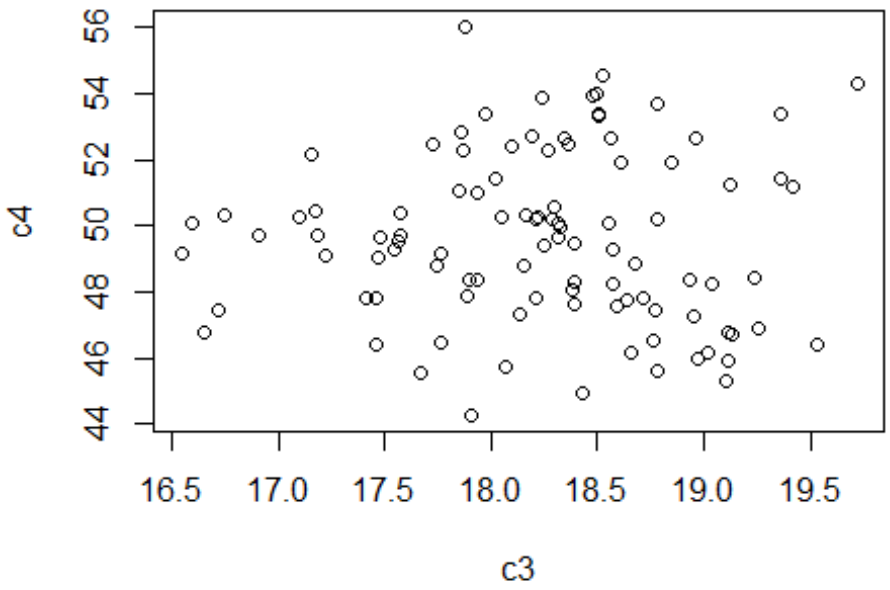
```
plot(density(c4))
```



Часть 5. Классификация неоднородных данных

Построить диаграмму рассеяния (plot) для двух переменных, которые имеют наименьшую корреляцию (по матрице корреляции из 3 части). Переменная с меньшим номером должна находится по оси Ox.

```
plot(c3,c4)
```



Объединить выбранные переменные в одну матрицу (cbind) и сохранить ее в новую переменную. Вывести первые три строки полученной матрицы (head).

```
data<-cbind(c3,c4)
head(data,3)
```

```
##          c3    c4
## [1,] 16.644 46.77
## [2,] 18.210 47.82
## [3,] 19.536 46.44
```

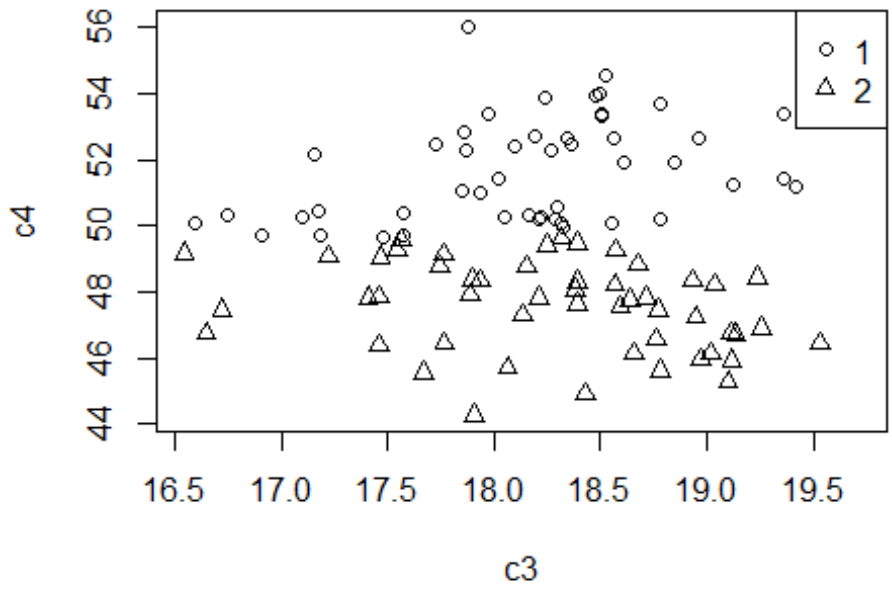
Выполнить кластерный анализ (kmeans) с разбиением на два класса в пространстве выбранных переменных (kmeans). В качестве первого аргумента функции kmeans необходимо передавать матрицу, а второго - количество классов (2), на которое производится разбиение.
Сохранить результаты в переменную kmres и вывести частоты наблюдений в каждом оцененном классе (table).

```
k = kmeans(data,2)
table(k$cluster)
```

```
##
##  1  2
## 49 51
```

Вывести график (plot) с легендой (legend), на котором должна быть обозначена классовая принадлежность каждого наблюдения различными символами (параметр pch).

```
plot(data,pch=ifelse(k$cluster==1,1,2), xlab="c3", ylab="c4")
legend("topright",legend=c("1","2"),pch=c(1,2))
```



Отсоединить таблицу данных qc от списка текущих переменных (работа с данными закончена).

```
detach(qc)
```