Лабораторная работа 3

Могильников Дмитрий

2022-11-19

Задание 1

Найдите датасет для проведения корреляционного анализа, в котором должно быть не менее трех переменных, с которыми можно работать. Кратко опишите датасет. К отчету приложите файл с датасетом (если он встроенный, то не нужно).

Загрузим датасет и выведем его через head()

```
options(width = 100)
library("MASS")
boston_df <- Boston

head(boston_df)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

Таблица с описанием для каждой переменной в проедставленном датасете:

Feature Variable	Description
CRIM	per capita crime rate by town.
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
NOX	nitrogen oxides concentration (parts per 10 million).
RM	average number of rooms per dwelling.
AGE	proportion of owner-occupied units built prior to 1940.
DIS	weighted mean of distances to five Boston employment centres.
RAD	index of accessibility to radial highways.
TAX	full-value property-tax rate per $10,000.
PTRATIO	pupil-teacher ratio by town.
BLACK	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
LSTAT	lower status of the population (percent).
MEDV	median value of owner-occupied homes in $1000s.

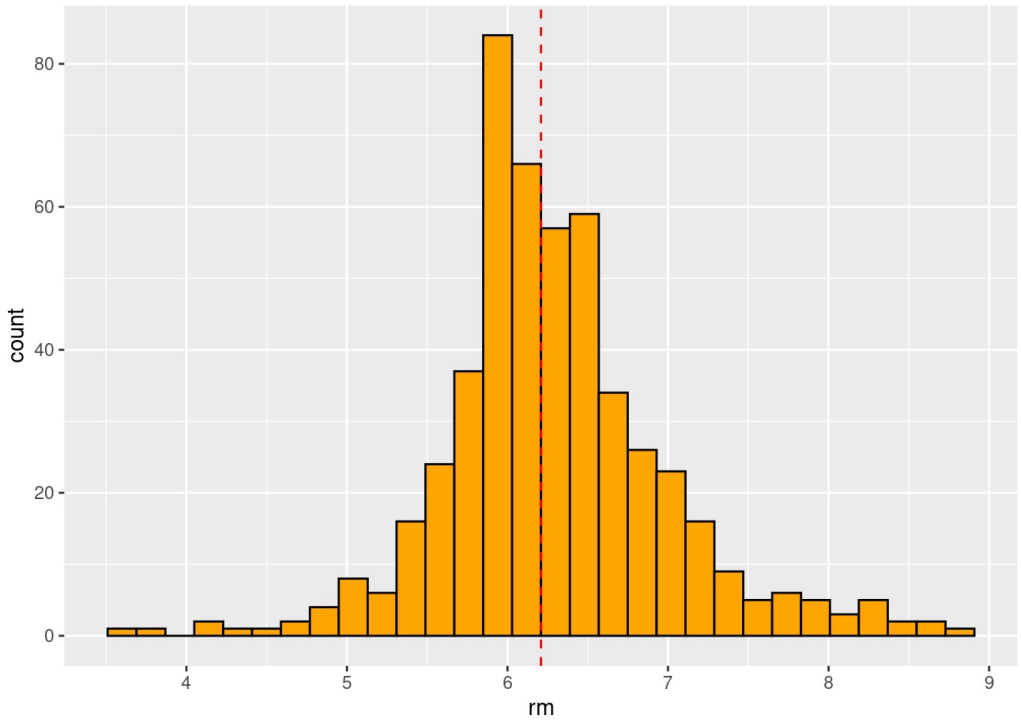
Задание 2

Проведите парный корреляционный анализ по двум переменным: постройте график рассеяния, найдите коэффициенты корреляции Пирсона, Спирмена и Кендала, а также проанализируйте статистическую значимость результатов(с помощью cor.test).

Парный корреляционный анализ будем проводить по следующим двум переменным: RM(количество комнат в доме) и MEDV(медианная стоимость домов)

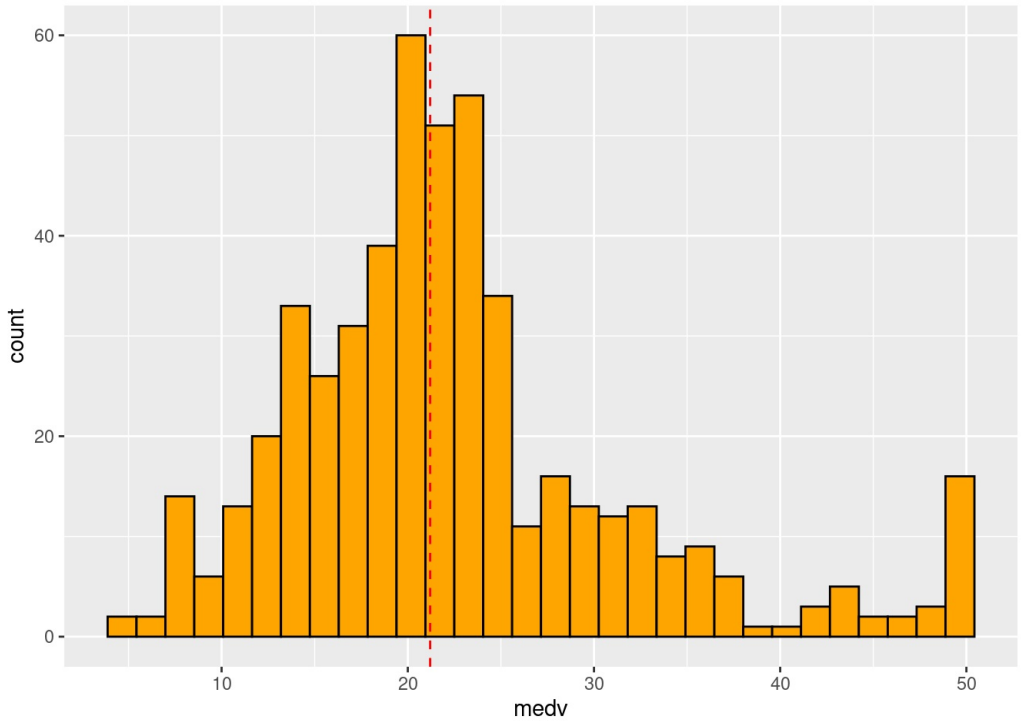
```
library(ggplot2)
attach(boston_df)
# Посмотрим на законы распределения этих переменых
ggplot(data = boston_df, aes(x = rm)) +
  geom_histogram(fill = "orange",
                 color = "black") +
  geom_vline(xintercept = median(rm),
             color = "red",
             lty = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



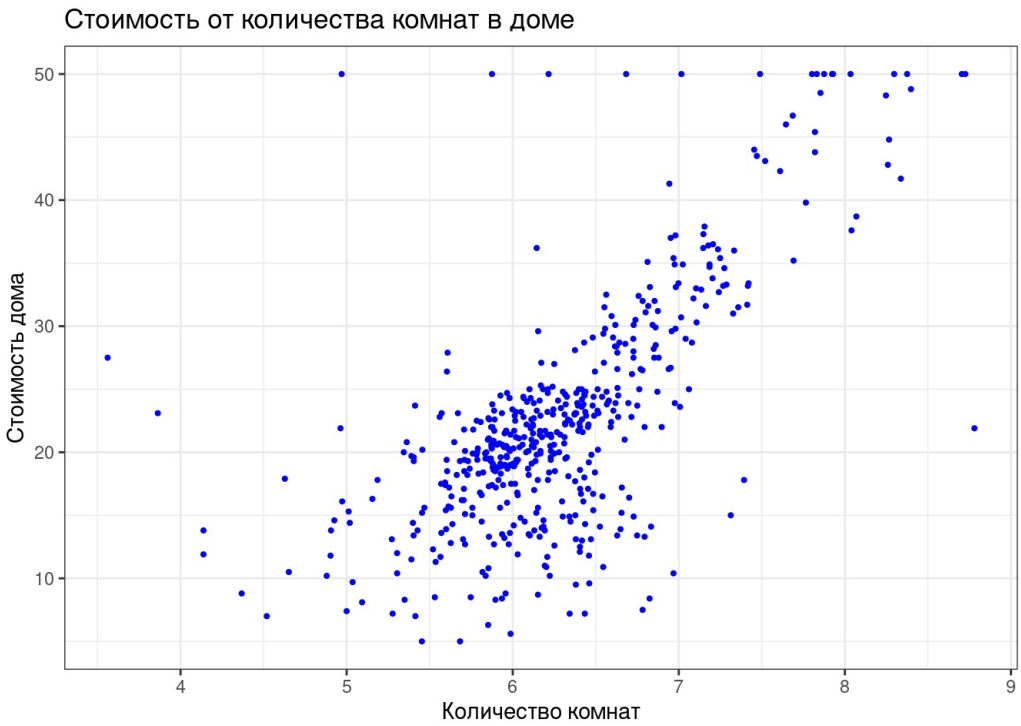

```
ggplot(data = boston_df, aes(x = medv)) +
  geom_histogram(fill = "orange",
                 color = "black") +
  geom_vline(xintercept = median(medv),
             color = "red",
             lty = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# Распределения близки к нормальным

#Построим график рассеяния:
ggplot(data = boston_df, aes(x = rm, y = medv)) +
  geom_point(size=0.75, color="blue") +
  labs(title="Стоимость от количества комнат в доме",
       x="Количество комнат",
       y="Стоимость дома") +
  theme_bw()
```




```
#Найдем коэффициенты корреляции и проанализируем статистическую значимость результатов
#Пирсон:
cor(x=rm, y=medv, method="pearson")
```

```
## [1] 0.6953599
```

```
cor.test(x=rm, y=medv, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  rm and medv
## t = 21.722, df = 504, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6474346 0.7378075
## sample estimates:
##       cor
## 0.6953599
```

```
#Спирман:
cor(x=rm, y=medv, method="spearman")
```

```
## [1] 0.6335764
```

```
cor.test(x=rm, y=medv, method = "spearman")
```

```
## Warning in cor.test.default(x = rm, y = medv, method = "spearman"): Есть совпадающие значения: не
## могу высчитать точное p-значение
```

```
##
##  Spearman's rank correlation rho
##
## data:  rm and medv
## S = 7911922, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.6335764
```

```
#Кендалл:
cor(x=rm, y=medv, method="kendall")
```

```
## [1] 0.4828293
```

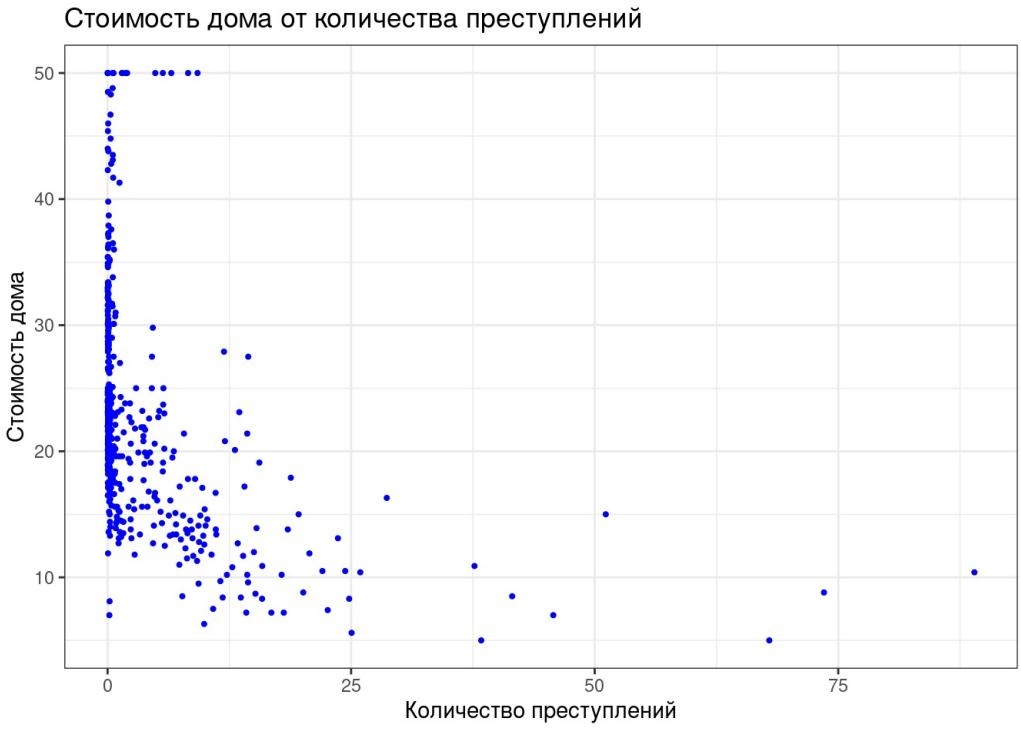
```
cor.test(x=rm, y=medv, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  rm and medv
## z = 16.192, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.4828293
```

```
#Исходя из полученных данных, можем сделать вывод, что наилучшая корреляция достигается с использованием коэффициента Пирсона. Как видим, рассчитанные коэффициенты корреляции статистичетки значимо отличаются от нуля (p-value < 2.2e-16)
```

Рассмотрим корреляцию стоимости с еще одной переменной: CRIM(количество преступлений на душу населения) и MEDV(медианная стоимость домов)

```
#Будем рассматривать на примере коэффициента корреляции Пирсона.
ggplot(data = boston_df, aes(x = crim, y = medv)) +
  geom_point(size=0.75, color="blue") +
  labs(title="Стоимость дома от количества преступлений",
       x="Количество преступлений",
       y="Стоимость дома") +
  theme_bw()
```

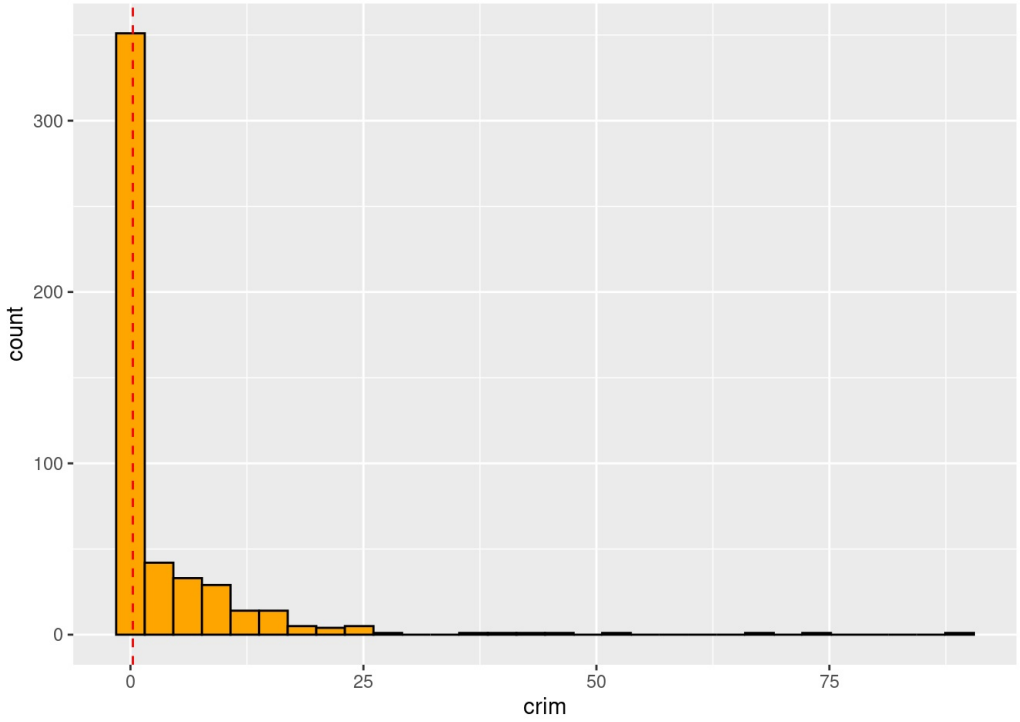



```
cor.test(crim, medv)
```

```
##
##  Pearson's product-moment correlation
##
## data:  crim and medv
## t = -9.4597, df = 504, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4599064 -0.3116859
## sample estimates:
##        cor
## -0.3883046
```

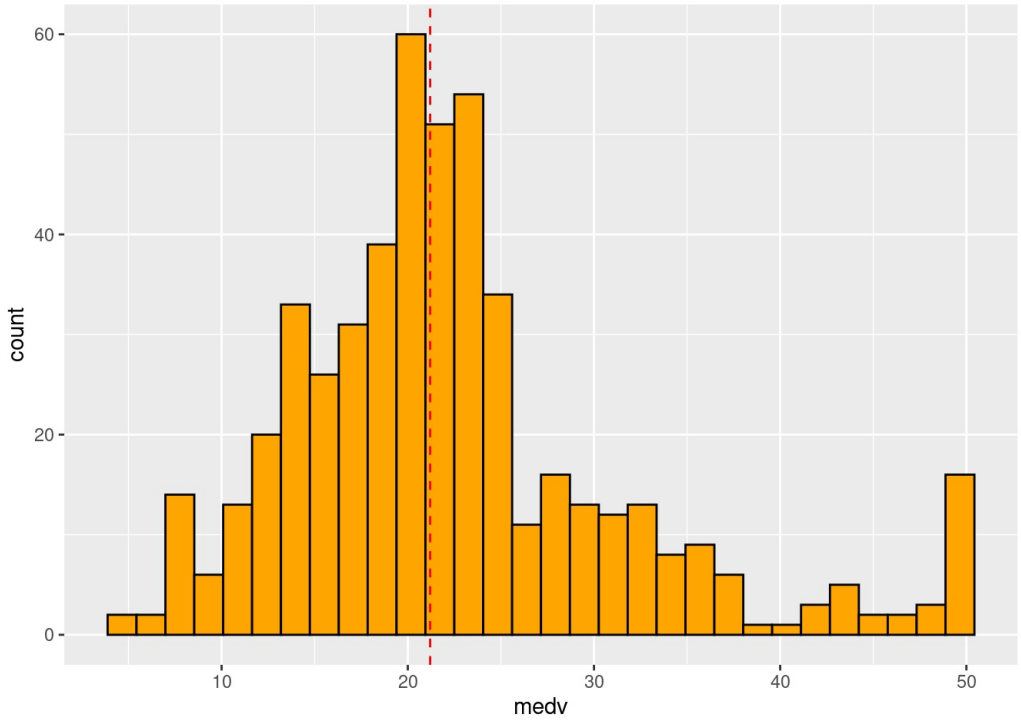
```
#Коэффициент кореляции между этими переменными не очень большой. Посмотрим на законы распределения этих переменых
ggplot(data = boston_df, aes(x = crim)) +
  geom_histogram(fill = "orange",
                 color = "black") +
  geom_vline(xintercept = median(crim),
             color = "red",
             lty = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



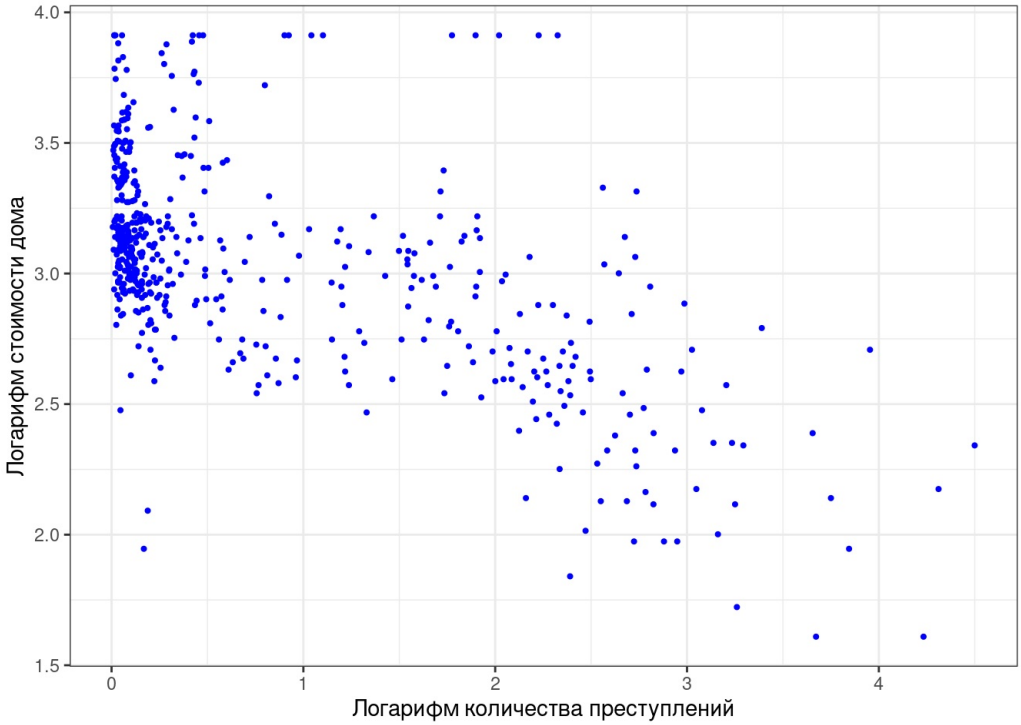

```
ggplot(data = boston_df, aes(x = medv)) +
  geom_histogram(fill = "orange",
                 color = "black") +
  geom_vline(xintercept = median(medv),
             color = "red",
             lty = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Видим, что гистограмма преступлений не подчиняется нормальному распределению. Прологарифмируем эти данные и снов
а посмотрим на коэффициент корреляции
crim_ln = log(crim+1)
medv_ln = log(medv)

ggplot(data = boston_df, aes(x = crim_ln, y = medv_ln)) +
  geom_point(size=0.75, color="blue") +
  labs(x="Логарифм количества преступлений",
       y="Логарифм стоимости дома") +
  theme_bw()
```




```
cor.test(crim_ln, medv_ln)
```

```
##
##  Pearson's product-moment correlation
##
## data:  crim_ln and medv_ln
## t = -16.828, df = 504, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6528186 -0.5408908
## sample estimates:
##        cor
## -0.5997806
```

```
#После преобразования данных коэффициент корреляции стал выше почти в два раза.
```

Задание 3

Постройте корреляционную матрицу для всех переменных датасета (минимально трех).

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(boston_df),
         method = "color",
         addCoef.col = 1,
         type="upper",
         number.cex = 0.6)
```