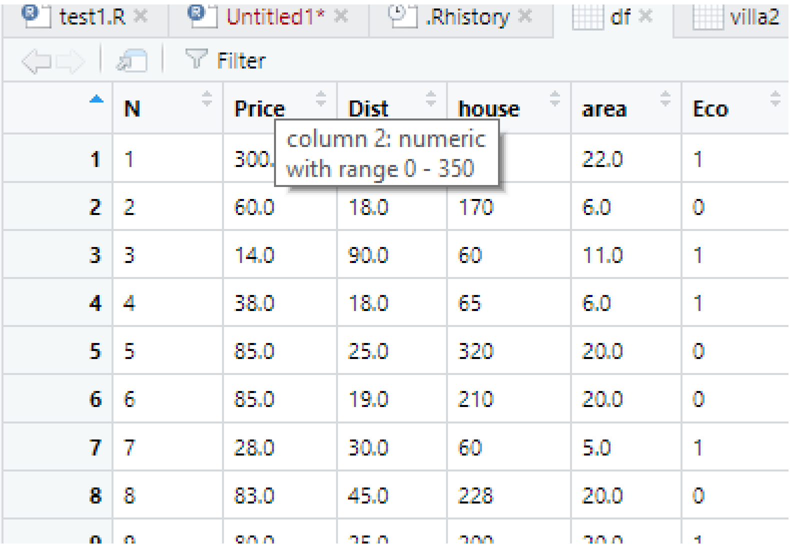
Задания

Задача 1:

Работа программы:

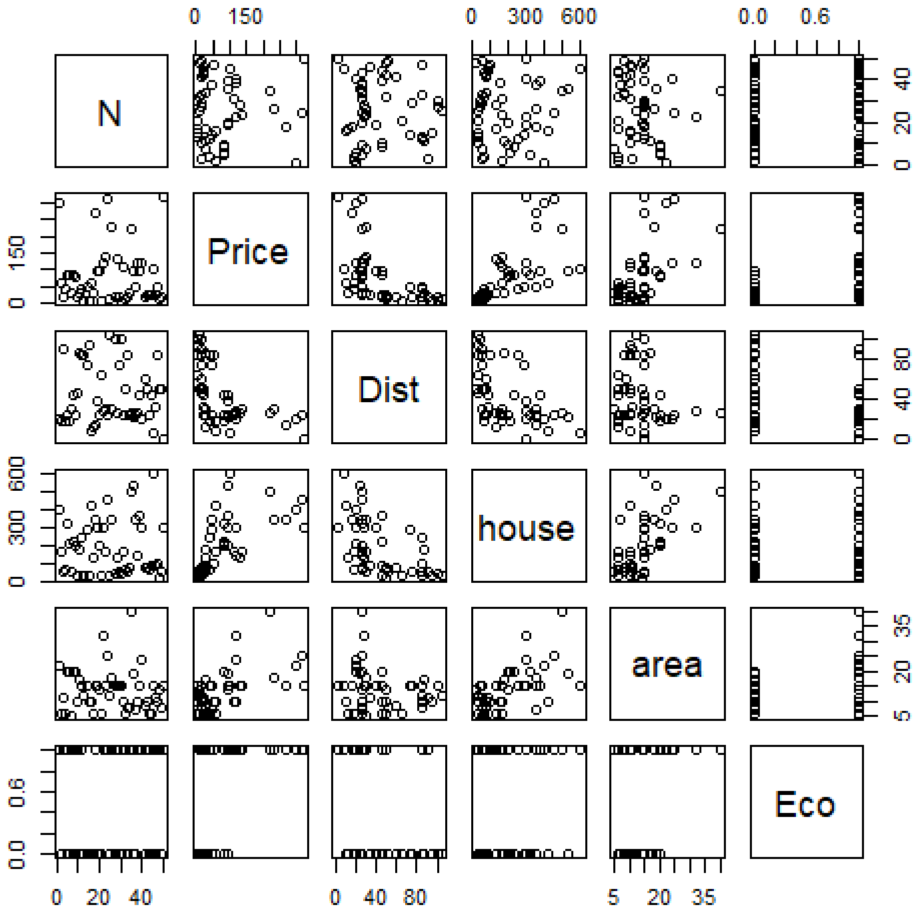
1



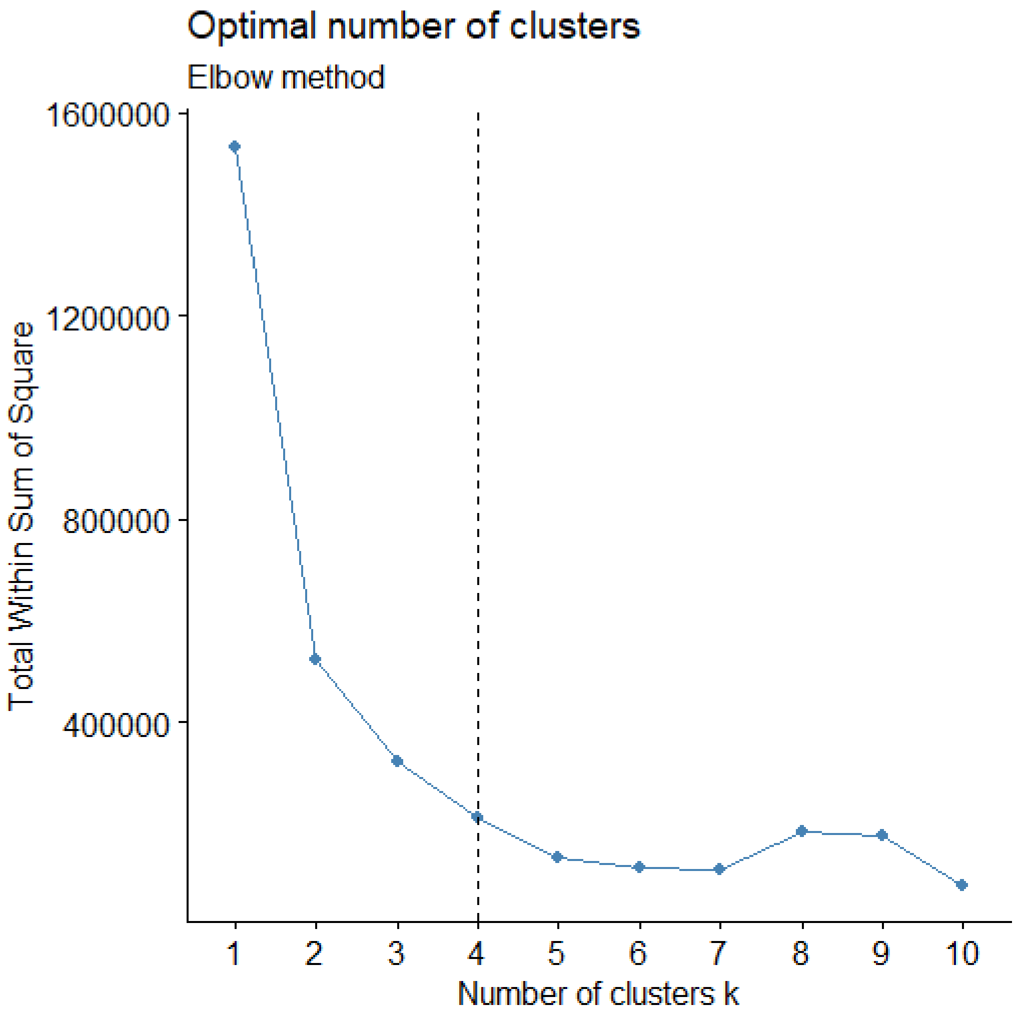
	N	Price	Dist	house	area	Eco
1	1	300.			22.0	1
2	2	60.0	18.0	170	6.0	0
3	3	14.0	90.0	60	11.0	1
4	4	38.0	18.0	65	6.0	1
5	5	85.0	25.0	320	20.0	0
6	6	85.0	19.0	210	20.0	0
7	7	28.0	30.0	60	5.0	1
8	8	83.0	45.0	228	20.0	0
9	9	80.0	25.0	200	20.0	1

column 2: numeric with range 0 - 350

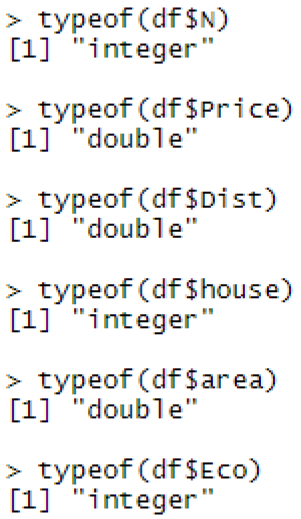
2



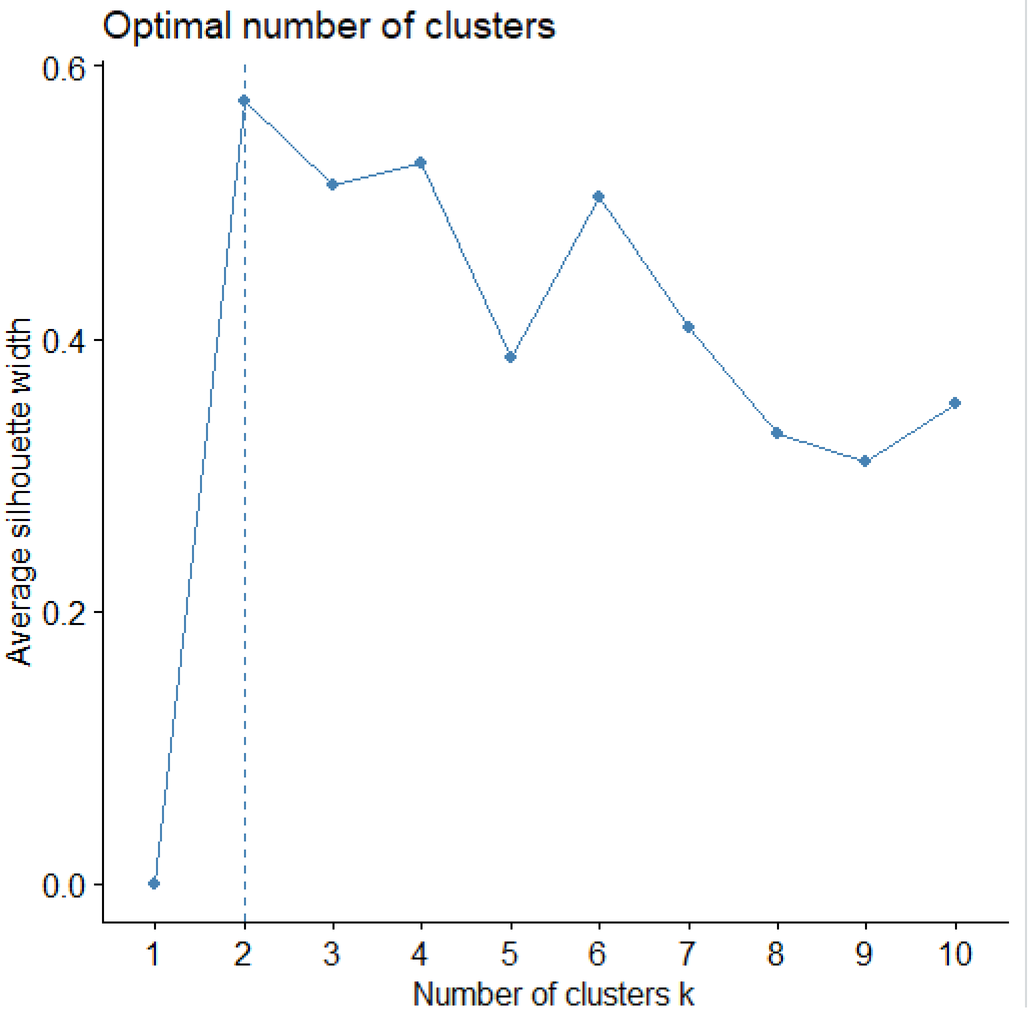
3

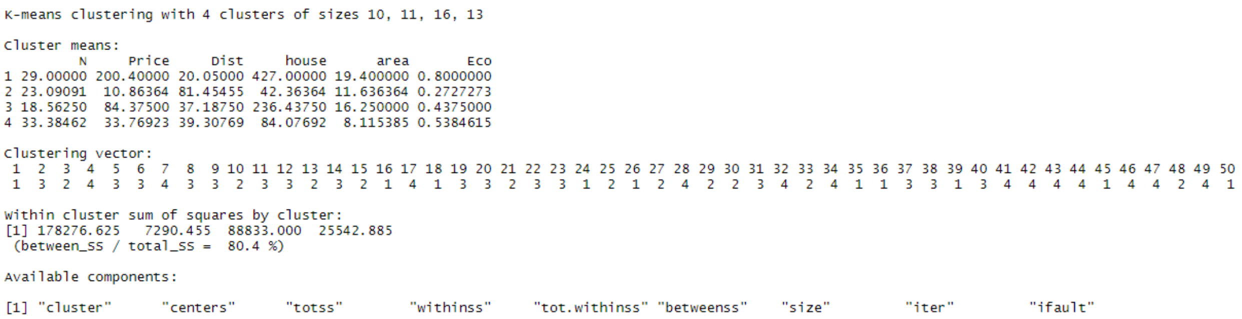


```
> dim(df)
[1] 50  6

> str(df)
'data.frame':   50 obs. of  6 variables:
 $ N    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Price: num  300 60 14 38 85 85 28 83 80 15 ...
 $ Dist : num  20 18 90 18 25 19 30 45 25 46 ...
 $ house: int  400 170 60 65 320 210 60 228 200 36 ...
 $ area : num  22 6 11 6 20 20 5 20 20 10 ...
 $ Eco  : int  1 0 1 1 0 0 1 0 1 1 ...
```

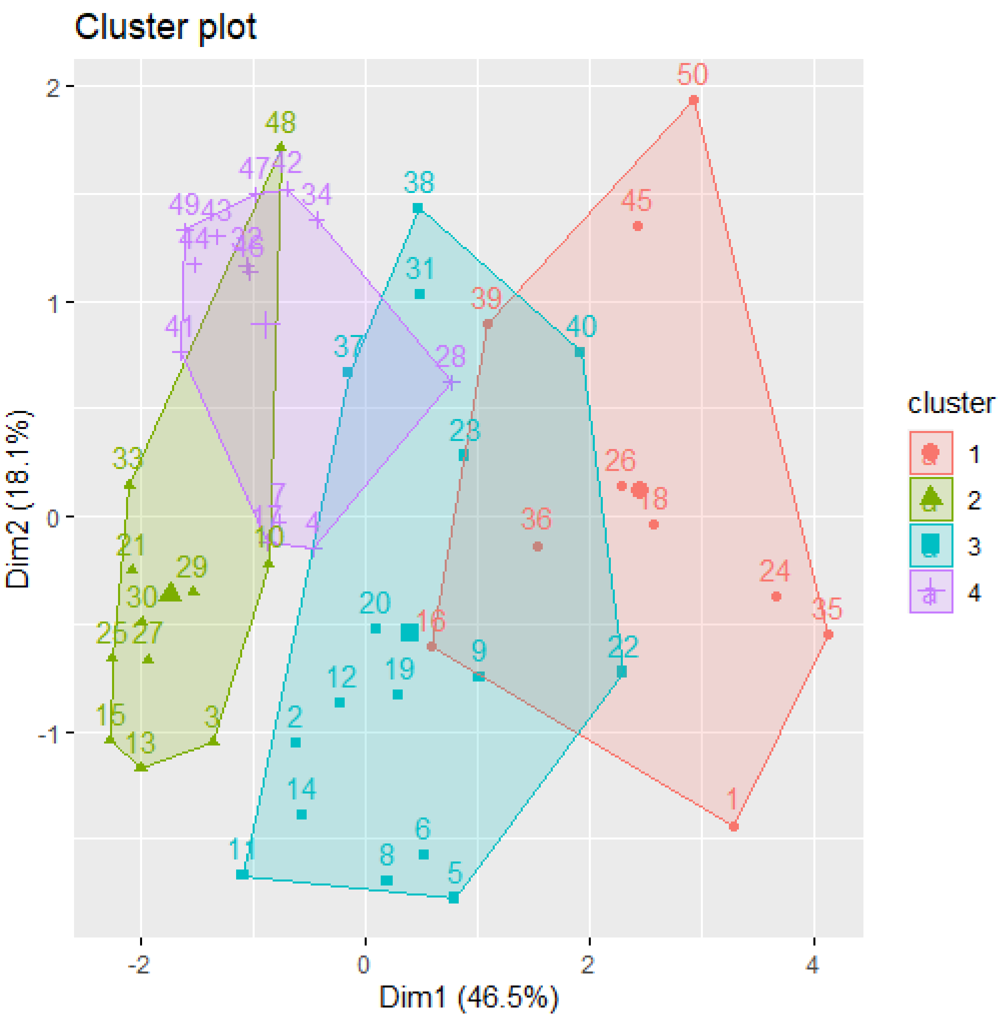


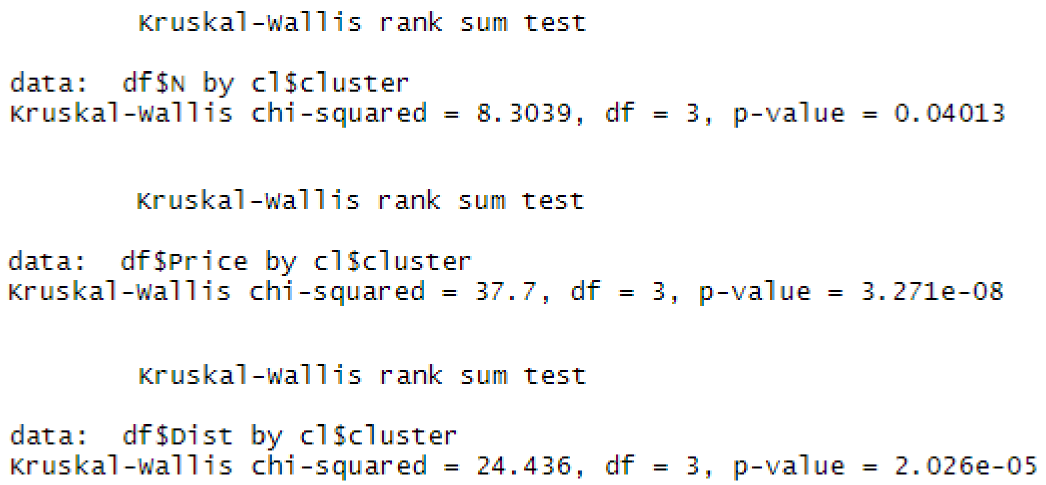
```
> typeof(df$N)
[1] "integer"

> typeof(df$Price)
[1] "double"

> typeof(df$Dist)
[1] "double"

> typeof(df$house)
[1] "integer"

> typeof(df$area)
[1] "double"

> typeof(df$Eco)
[1] "integer"
```



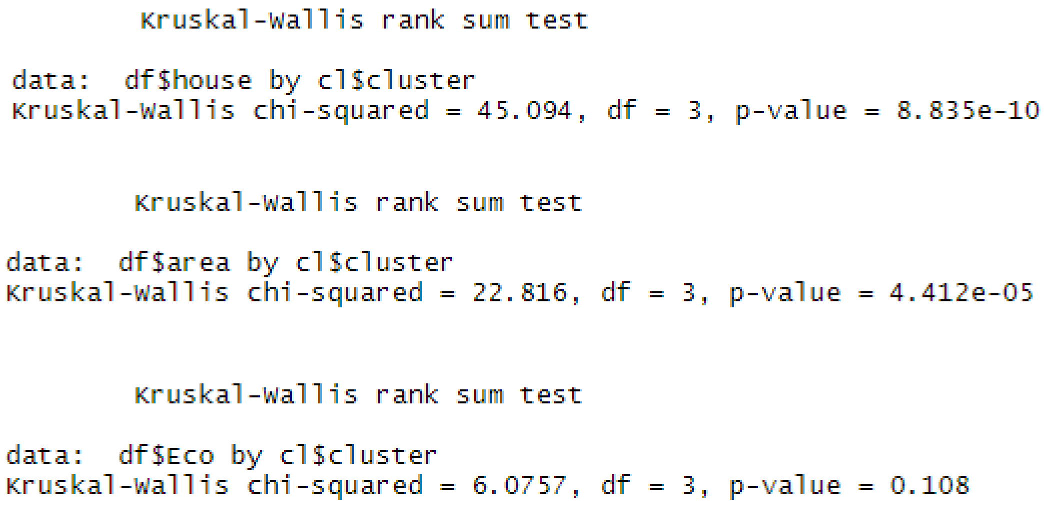


```
K-means clustering with 4 clusters of sizes 10, 11, 16, 13

Cluster means:
         N     Price     Dist     house      area       Eco
1 29.00000 200.40000 20.05000 427.00000 19.400000 0.8000000
2 23.09091  10.86364 81.45455  42.36364 11.636364 0.2727273
3 18.56250  84.37500 37.18750 236.43750 16.250000 0.4375000
4 33.38462  33.76923 39.30769  84.07692  8.115385 0.5384615

Clustering vector:
 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 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 
 1  3  2  4  3  3  4  3  3  2  3  3  2  3  2  1  4  1  3  3  2  3  3  1  2  1  2  4  2  2  3  4  2  4  1  1  3  3  1  3  4  4  4  4  1  4  4  2  4  1 

Within cluster sum of squares by cluster:
[1] 178276.625   7290.455  88833.000  25542.885
 (between_SS / total_SS =  80.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"      
```





```
        Kruskal-Wallis rank sum test

data:  df$N by cl$cluster
Kruskal-Wallis chi-squared = 8.3039, df = 3, p-value = 0.04013


        Kruskal-Wallis rank sum test

data:  df$Price by cl$cluster
Kruskal-Wallis chi-squared = 37.7, df = 3, p-value = 3.271e-08


        Kruskal-Wallis rank sum test

data:  df$Dist by cl$cluster
Kruskal-Wallis chi-squared = 24.436, df = 3, p-value = 2.026e-05
```



```
        Kruskal-Wallis rank sum test

data:  df$house by cl$cluster
Kruskal-Wallis chi-squared = 45.094, df = 3, p-value = 8.835e-10


        Kruskal-Wallis rank sum test

data:  df$area by cl$cluster
Kruskal-Wallis chi-squared = 22.816, df = 3, p-value = 4.412e-05


        Kruskal-Wallis rank sum test

data:  df$Eco by cl$cluster
Kruskal-Wallis chi-squared = 6.0757, df = 3, p-value = 0.108
```

Гипотезу о различии распределения по кластерам отвергаем, тк p-value близко к 0

Листинг:

```
#install.packages(c("factoextra"))
library(readr)
library(factoextra)

df <- read_delim("villa2.csv", ";", escape_double = FALSE, locale =
locale(decimal_mark = ","), trim_ws = TRUE)
View(df)
plot(df)
df <- read.csv(file = "villa2.csv", sep = ";")
st <- as.vector(df$area);n <- gsub(",", ".", st);n <- as.numeric(n);df$area <- n
st <- as.vector(df$Price);n <- gsub(",", ".", st);n <- as.numeric(n);df$Price <- n
st <- as.vector(df$Dist);n <- gsub(",", ".", st);n <- as.numeric(n);df$Dist <- n
dim(df)
str(df)
typeof(df$N)
typeof(df$Price)
typeof(df$Dist)
typeof(df$house)
typeof(df$area)
typeof(df$Eco)
df <- na.omit(df)
fviz_nbclust(df, kmeans, method = "wss") + labs(subtitle = "Elbow method") + geom_vli
ne(xintercept = 4, linetype = 2)
fviz_nbclust(df, kmeans, method = "silhouette")
fviz_cluster(kmeans(df, 4), data = df, ellipse.type = 'convex')
kruskal.test(df$N ~ cl$cluster)
kruskal.test(df$Price ~ cl$cluster)
kruskal.test(df$Dist ~ cl$cluster)
kruskal.test(df$house ~ cl$cluster)
kruskal.test(df$area ~ cl$cluster)
kruskal.test(df$Eco ~ cl$cluster)
```