Лабораторная работа №4

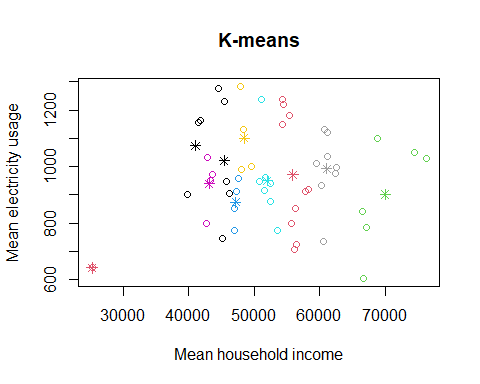
Голуб Светлана

# K-means

**Метод K-средних** (k-means) - метод кластеризации, алгоритм которого стремиться минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

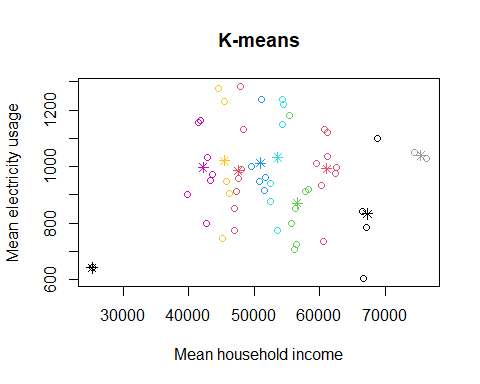
1. Объединяем штаты в кластеры и создаём диаграмму кластеров:

load("D:/MyDoc/Desktop/МИФЯГА/data\_science/lab4/income\_elec\_state.Rdata")  
income <- income\_elec\_state$income  
elec <- income\_elec\_state$elec  
test <- scale(income\_elec\_state)  
income\_elec\_state.km <- kmeans(income\_elec\_state, 10)  
income\_elec\_state.cl <- income\_elec\_state.km$cluster  
plot(income, elec, col = income\_elec\_state.cl, xlab="Mean household income", ylab="Mean electricity usage", main="K-means")  
points(income\_elec\_state.km$centers, col=1:10, pch=8)



1. При повторном выполнении кластеризации, мы можем получить другое разбиение. Это происходит из-за того, что на первом шаге алгоритма выбираются произвольные центры, а количество итераций алгоритма по умолчанию равно 10, что иногда прерывает алгоритм до того, как он найдёт наилучшее разбиение.  
   Для того, чтобы снизить вероятность изменения разбиения, можно увеличить значение параметра nstart (количество запусков алгоритма на случайных числах) и увеличит значение iter.max (максимальное количество иттераций)

income\_elec\_state.km <- kmeans(income\_elec\_state, 10, nstart = 20, iter.max = 20)  
income\_elec\_state.cl <- income\_elec\_state.km$cluster  
plot(income, elec, col = income\_elec\_state.cl, xlab="Mean household income", ylab="Mean electricity usage", main="K-means")  
points(income\_elec\_state.km$centers, col=1:10, pch=8)



Теперь при каждом следующем запуске алгоритма разбиение на кластеры одинаковое

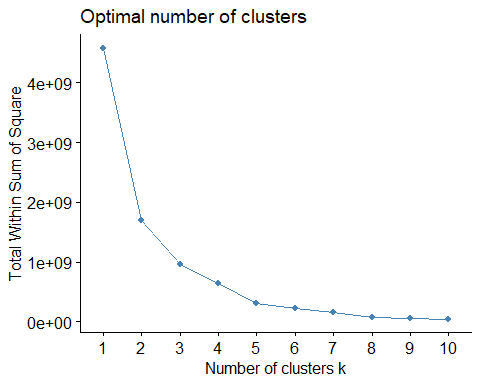
с. По диаграмме кластеров можно заметить, что разбиение не является оптимальным. Для того, чтобы найти наиболее подходящее число кластеров для разбиения, воспользуемся методом “локтя”.

library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

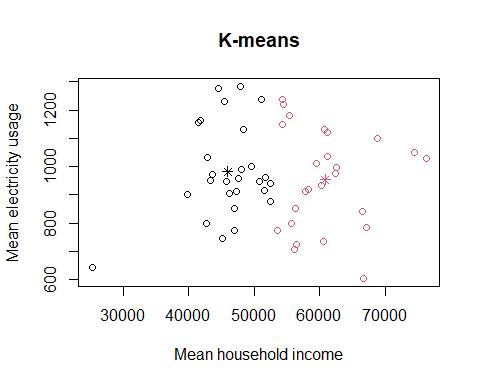
fviz\_nbclust(income\_elec\_state, kmeans, method = "wss")



На графике виден чёткий перелом в районе значений 2-5.

* Разделим на 2 кластера

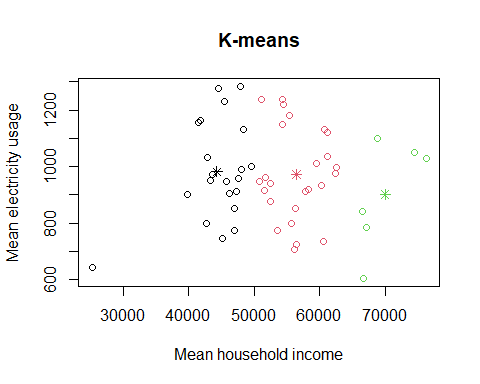
income\_elec\_state.km <- kmeans(income\_elec\_state, 2, nstart = 20, iter.max = 20)  
income\_elec\_state.cl <- income\_elec\_state.km$cluster  
plot(income, elec, col = income\_elec\_state.cl, xlab="Mean household income", ylab="Mean electricity usage", main="K-means")  
points(income\_elec\_state.km$centers, col=1:10, pch=8)



При таком разделении точка-выброс включена в один из общих кластеров. Такого быть не должно.

* Разделим на 3 кластера

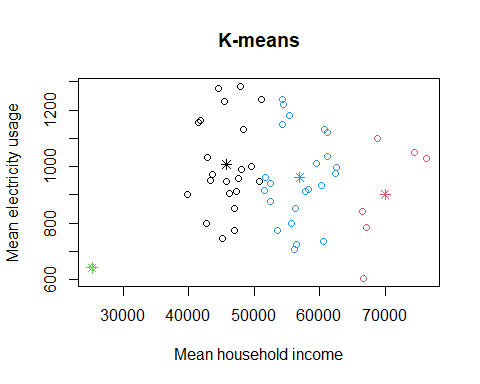
income\_elec\_state.km <- kmeans(income\_elec\_state, 3, nstart = 20, iter.max = 20)  
income\_elec\_state.cl <- income\_elec\_state.km$cluster  
plot(income, elec, col = income\_elec\_state.cl, xlab="Mean household income", ylab="Mean electricity usage", main="K-means")  
points(income\_elec\_state.km$centers, col=1:10, pch=8)



Проблема осталась та же.

-Разделим на 4 кластера

income\_elec\_state.km <- kmeans(income\_elec\_state, 4, nstart = 20, iter.max = 20)  
income\_elec\_state.cl <- income\_elec\_state.km$cluster  
plot(income, elec, col = income\_elec\_state.cl, xlab="Mean household income", ylab="Mean electricity usage", main="K-means")  
points(income\_elec\_state.km$centers, col=1:10, pch=8)



Разделение выглядит оптимальным.

с. Преобразуем наши данные в логарифмический мастаб и посмотрим, изменилась ли диаграмма кластеризации:

income\_elec <- income\_elec\_state  
income\_elec[,c(1,2)] <- log10(income\_elec[,c(1,2)])  
income\_elec.km <- kmeans(income\_elec, 5, nstart = 20, iter.max = 20)  
income\_elec.cl <- income\_elec.km$cluster  
plot(income\_elec$income, income\_elec$elec, col = income\_elec.cl, xlab="Mean household income", ylab="Mean electricity usage", main="K-means for log")  
points(income\_elec.km$centers, col=1:10, pch=8)

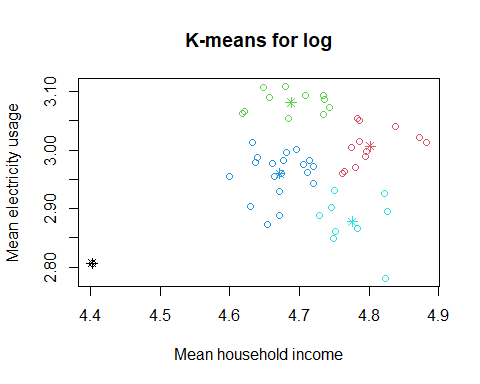
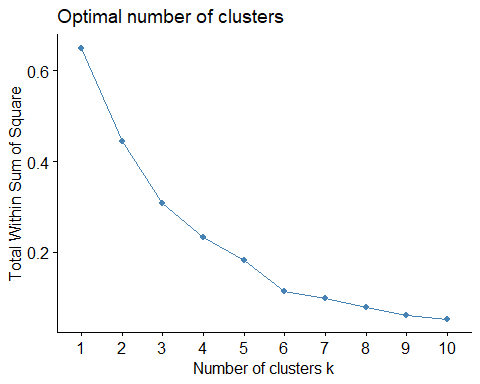


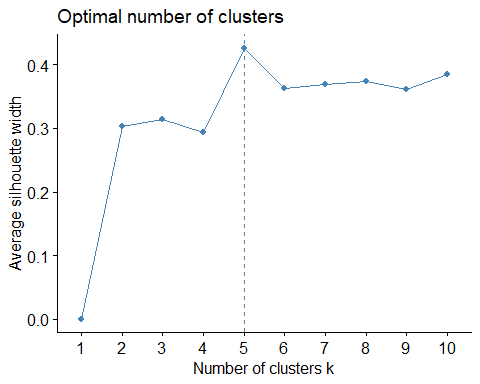
Диаграмма изменилась, так как алгоритм k-means чувствителен к любому мастабированию. Результат получится иным, даже если мы одинаково трансформируем все данные.

1. Посмотрим, как изменилось оптимальное колисетво кластеров:

fviz\_nbclust(income\_elec, kmeans, method = "wss")

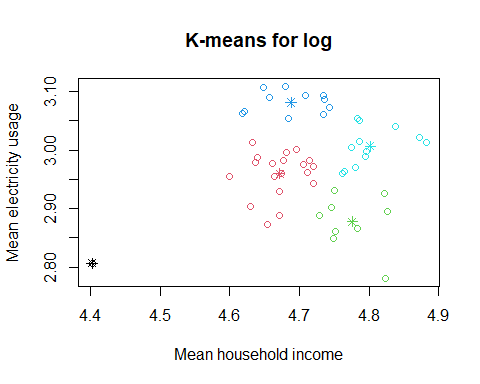
 Не совсем понятно, какое количество кластеров лучше подходит. Попробуем использовать другой метод:

fviz\_nbclust(income\_elec, kmeans, method = "silhouette")



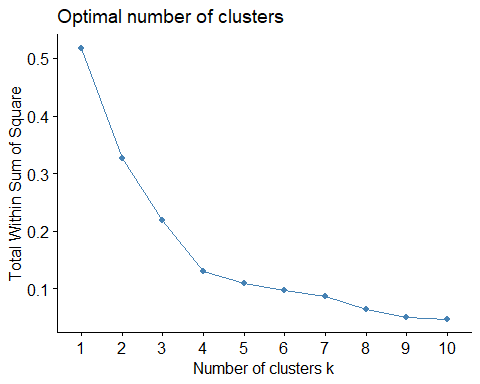
Видим, что 5 кластеров наиболее оптимальное количество. Строим диаграмму кластеров:

income\_elec.km <- kmeans(income\_elec, 5, nstart = 20, iter.max = 20)  
income\_elec.cl <- income\_elec.km$cluster  
plot(income\_elec$income, income\_elec$elec, col = income\_elec.cl, xlab="Mean household income", ylab="Mean electricity usage", main="K-means for log")  
points(income\_elec.km$centers, col=1:10, pch=8)

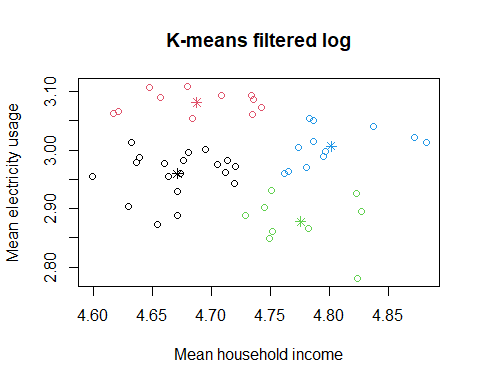


1. В наших данных легко можно заметить выброс - крайняя левая точка. Попробуем убрать Пуэрто-Рико и пересмотреть количество кластеров:

income\_elec <- subset(income\_elec, income\_elec$income > 4.5)  
fviz\_nbclust(income\_elec, kmeans, method = "wss")

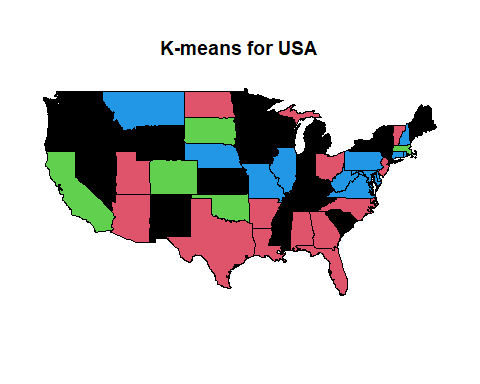
 По графику видно, что оптимальное значение количества кластеров - 4.

income\_elec.km <- kmeans(income\_elec, 4, nstart = 10, iter.max = 10)  
income\_elec.cl <- income\_elec.km$cluster  
plot(income\_elec$income, income\_elec$elec, col = income\_elec.cl, xlab="Mean household income", ylab="Mean electricity usage", , main="K-means filtered log")  
points(income\_elec.km$centers, col=1:10, pch=8)



1. Изобразим принадлежности штатов к кластерам на карте США:

library(maps)  
map\_order <- c('AL', 'AZ', 'AR', 'CA', 'CO', 'CT', 'DE', 'DC', 'FL',  
'GA', 'ID', 'IL', 'IN', 'IA', 'KS', 'KY', 'LA', 'ME',  
'MD', 'MA', 'MI', 'MN', 'MS', 'MO', 'MT', 'NE', 'NV',  
'NH', 'NJ', 'NM', 'NY', 'NC', 'ND', 'OH', 'OK', 'OR',  
'PA', 'RI', 'SC', 'SD', 'TN', 'TX', 'UT', 'VT', 'VA',  
'WA', 'WV', 'WI', 'WY')  
map\_color <- income\_elec.km$cluster[map\_order]  
map('state', col=map\_color, fill=TRUE)  
title("K-means for USA")



# Hierarchical clustering

**Иерархическая кластеризация** (Hierarchical clustering) - метод кластеризации, направленный на создание иерарии вложенных кластеров.

1. Для начала построим дендрограмму по исходным данным:

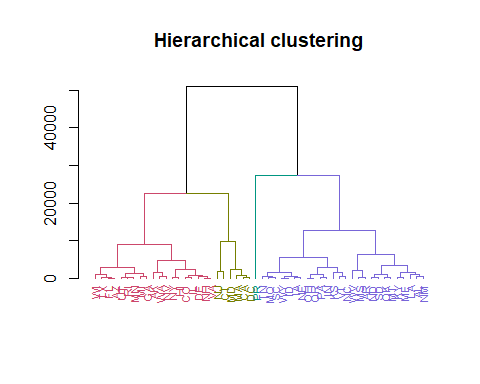
library(dendextend)

##   
## ---------------------  
## Welcome to dendextend version 1.14.0  
## Type citation('dendextend') for how to cite the package.  
##   
## Type browseVignettes(package = 'dendextend') for the package vignette.  
## The github page is: https://github.com/talgalili/dendextend/  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues  
## Or contact: <tal.galili@gmail.com>  
##   
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))  
## ---------------------

##   
## Attaching package: 'dendextend'

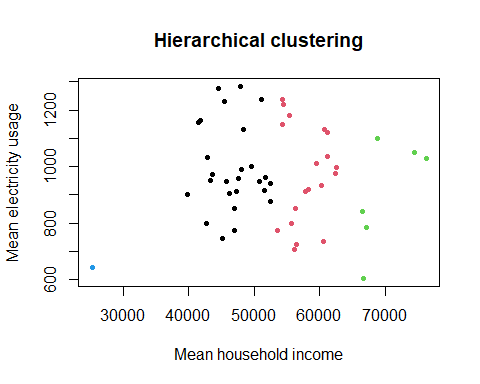
## The following object is masked from 'package:stats':  
##   
## cutree

hc <- hclust(dist(income\_elec\_state))  
tree <- as.dendrogram(hc)  
tree <- color\_branches(tree, k = 4)  
tree %>% set("labels\_col", k=4) %>% set("labels\_cex", 0.7) %>% plot(main="Hierarchical clustering")



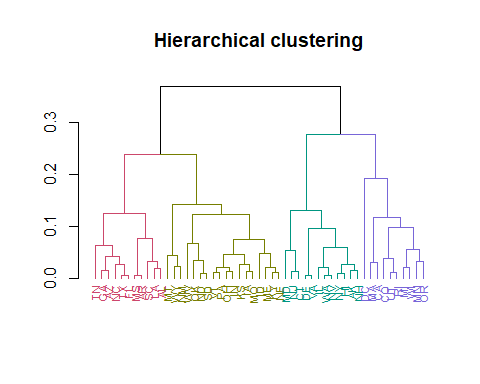
На диаграмме кластеров можно увидеть, как были объеденены штаты:

sub\_grp <- cutree(hc, k=4)  
plot(income\_elec\_state, pch=20, col=sub\_grp, main="Hierarchical clustering", xlab="Mean household income", ylab="Mean electricity usage")



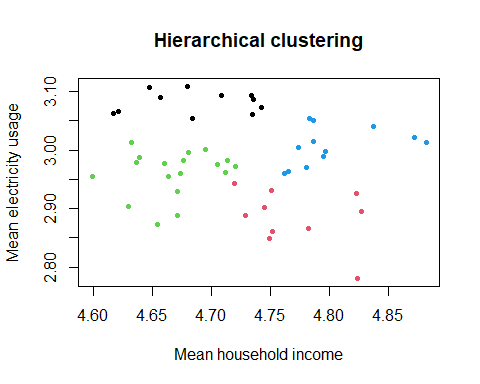
1. Теперь построим дендрограмму на основе отчищенных данных, приведённых к логорифмическому масштабу:

hc\_new <- hclust(dist(income\_elec))  
tree <- as.dendrogram(hc\_new)  
tree <- color\_branches(tree, k = 4)  
tree %>% set("labels\_col", k=4) %>% set("labels\_cex", 0.7) %>% plot(main = "Hierarchical clustering")



На диаграмме кластеров можно увидеть, как были объеденены штаты:

sub\_grp\_new <- cutree(hc\_new, k=4)  
plot(income\_elec, pch=20, col=sub\_grp\_new, main = "Hierarchical clustering", xlab="Mean household income", ylab="Mean electricity usage")



1. Изобразим результат иерархической кластеризации на карте США:

map\_color <- sub\_grp\_new[map\_order]  
map('state', col=map\_color, fill=TRUE)  
title("Hierarchical clustering")

