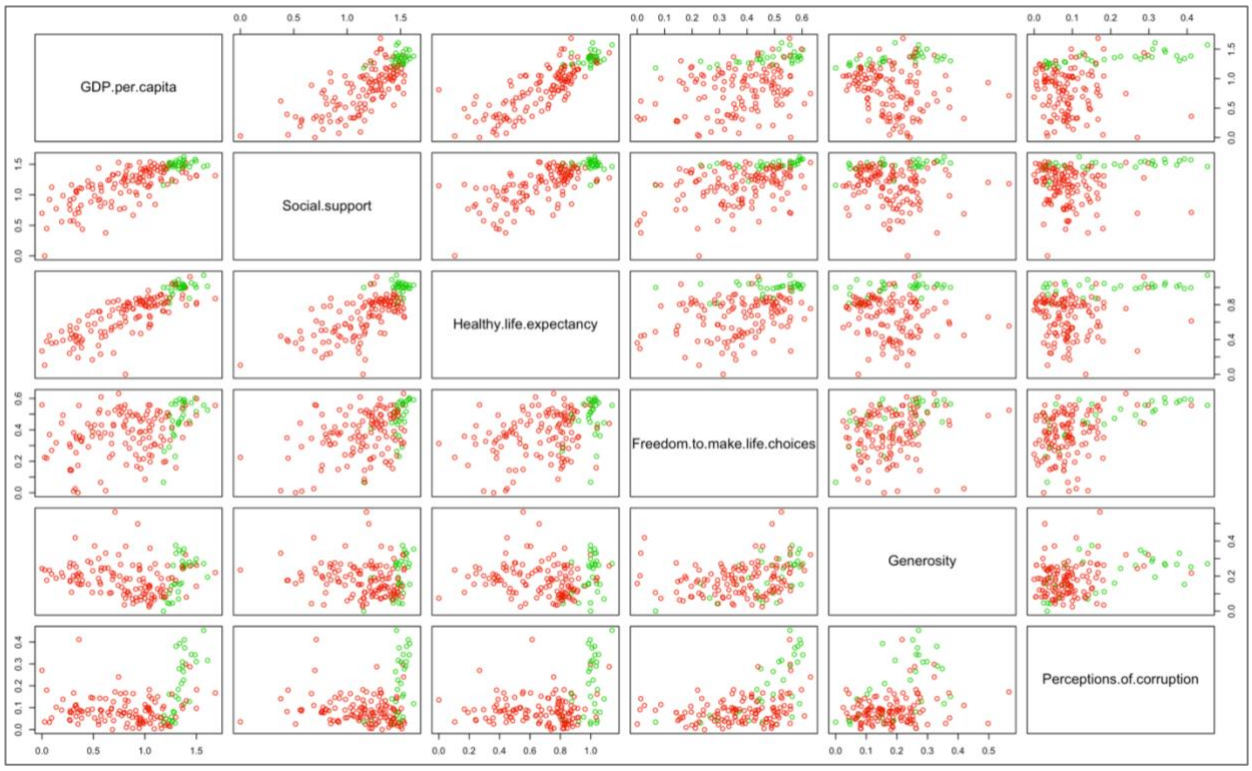
Домашнее задание №1 (Отчёт)

Пункт 1

Мы имеем выборку из 156 стран, 20% из которой – представители развитых стран. Для каждой страны присутствует информация о ВВП на душу населения, показателях социальной поддержки, ожидания здоровой жизни, свободы делать жизненный выбор, щедрости и ощущения коррумпированности. Данные переменные используется при подсчёте уровня счастья населения страны. Ниже представлены описательные статистики переменных:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Score	156	5.407	1.113	2.853	4.545	6.184	7.769
GDP.per.capita	156	0.905	0.398	0.000	0.603	1.233	1.684
Social.support	156	1.209	0.299	0.000	1.056	1.452	1.624
Healthy.life.expectancy	156	0.725	0.242	0.000	0.548	0.882	1.141
Freedom.to.make.life.choices	156	0.393	0.143	0.000	0.308	0.507	0.631
Generosity	156	0.185	0.095	0.000	0.109	0.248	0.566
Perceptions.of.corruption	156	0.111	0.095	0.000	0.047	0.141	0.453
Advanced	156	0.199	0.400	0	0	0	1

Парные графики компонент индекса счастья можно наблюдать далее (зелёным цветом помечены развитые страны, красным – развивающиеся):



Пункт 2

По компонентам индекса счастья ближайшими соседями Польши являются: Панама, Чехия, Косово, Уругвай и Перу. Единственной развитой страной из них является Чехия.

Пункт 3

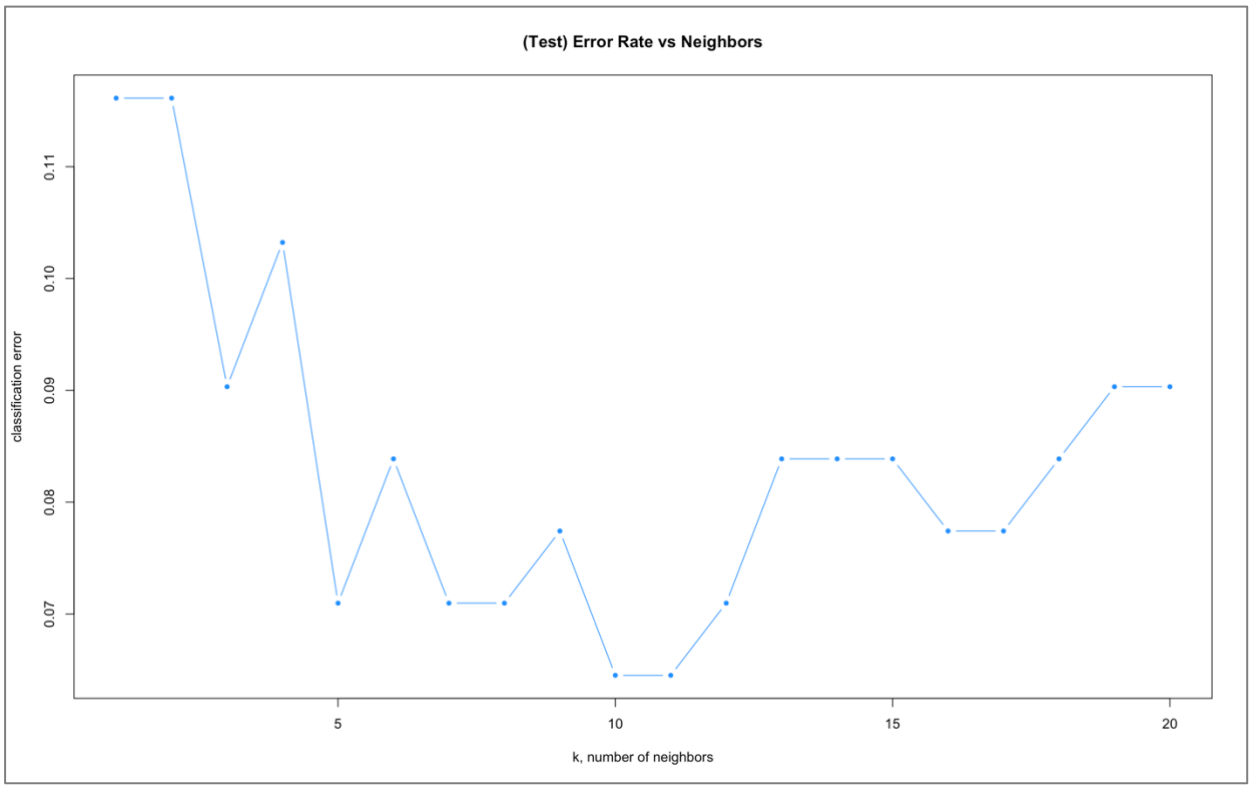
В одной из последних попыток прогона кода выиграл алгоритм kd-tree: 3 миллисекунды по сравнению с 4 в случае простого перебора. Правда более значительное преимущество алгоритм показывает при большом количестве соседей и б**о**льшем размере данных.

Пункт 4

Алгоритм, натренированный на компонентах индекса счастья 155 стран, предсказал Польше развивающийся класс, как и есть на самом деле. Это может быть свидетельством того, что уровень счастья в Польше соотносится с уровнем её экономического развития.

Пункт 5

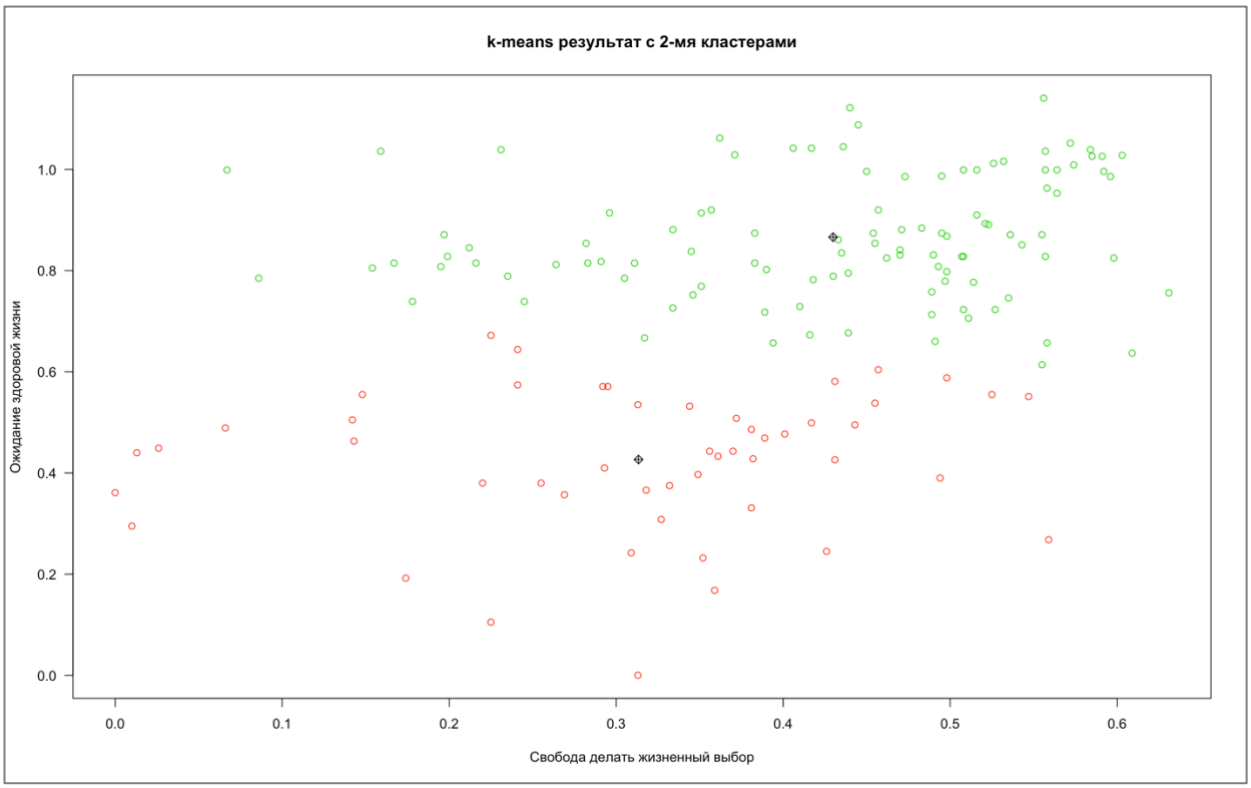
Алгоритм, натренированный на выборке без Польши, даёт ошибку в размере 0.07, что можно считать вполне неплохим результатом.



Исходя из графика, лучшим количеством соседей по критерию минимизации ошибки является число 10 или 11; чтобы алгоритму не пришлось рандомно выбирать класс в случае равенства соседей, остановимся на 11 соседях.

Пункт 6

В результате кластеризации центры кластеров получили данные координаты: (0.430, 0.866) и (0.313, 0.427). Таким образом, алгоритм попытался сделать разбивку по двум компонентам уровня счастья (с м**е**ньшей свободой выбора и ожидаемой длительностью жизни в одну группу, с б**о**льшей – в другую). Ближе всего к данным центрам были Мексика и Того соответственно. Польша попала в зелёную группу: с б**о**льшей свободой выбора и ожидаемой длительностью жизни. Графический результат кластеризации представлен ниже (чёрным цветом обозначены центры каждого из кластеров):

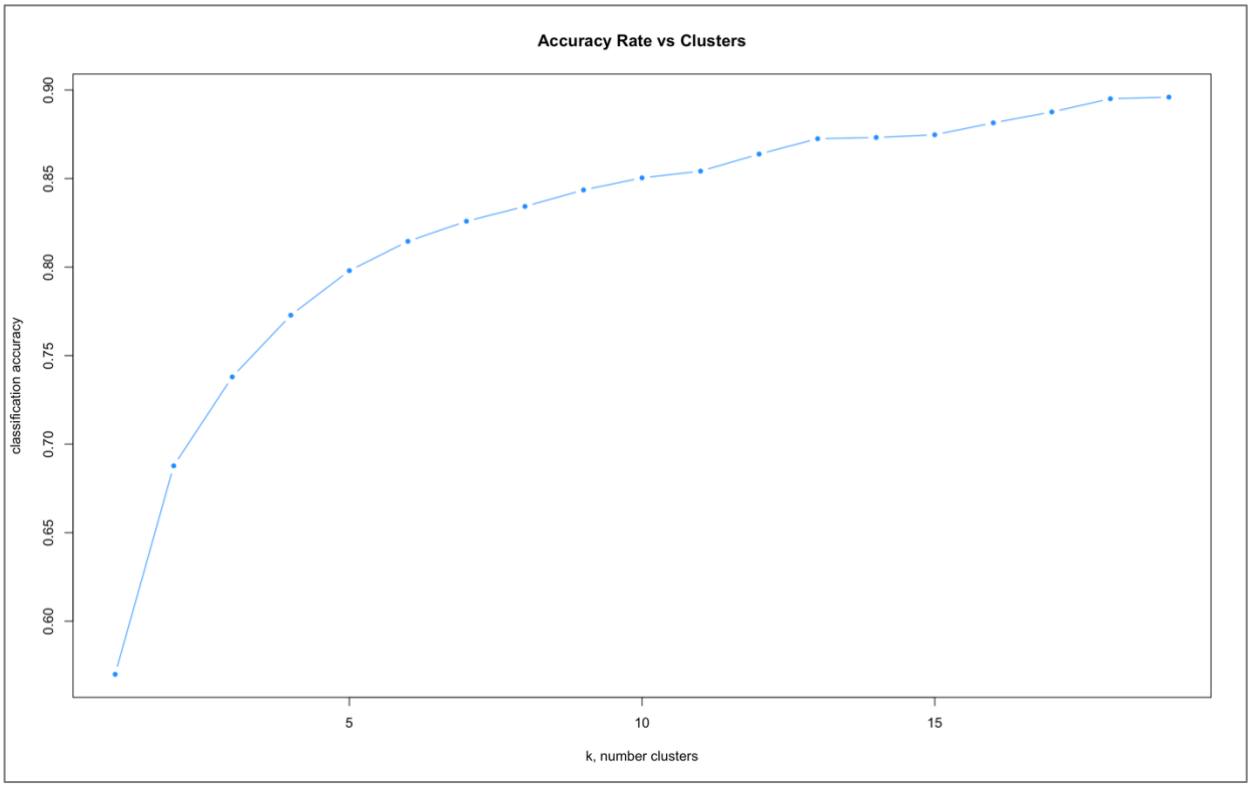


Пункт 7

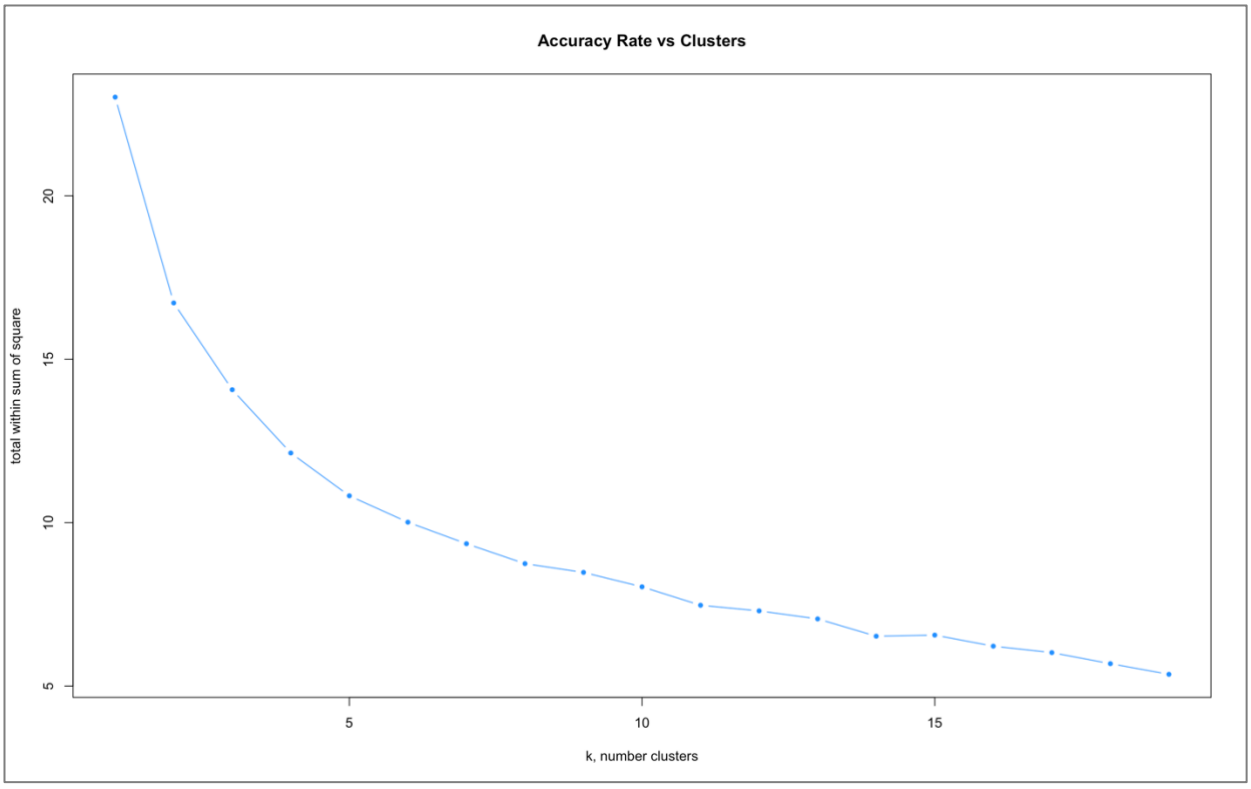
Польша оказалась в кластере с б**о**льшим ВВП на душу населения, социальной поддержкой и ожидаемой длительностью жизни, с б**о**льшей свободой выбора, правда с м**е**ньшим показателем щедрости и б**о**льшим ощущением коррупции.

Пункт 8

Для минимизации суммарного внутригруппового расстояния стоит выбирать как можно б**о**льшее количество кластеров, но это решение может расходиться с целями самой кластеризации. Поэтому стоит выбрать то количество кластеров, при котором увеличение кластеров на единичку не будет давать большого прироста в суммарном внутригрупповом расстоянии: в нашем случае стоит остановиться на 5 кластерах (мы построили отношение межгруппового расстояния к общему: общее не меняется, при увеличении межгруппового расстояния внутригрупповое расстояние снижается).



Также можно напрямую построить, как меняется суммарное внутригрупповое расстояние:



Получаем аналогичное относительно оптимальное решение – 5 кластеров.