МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МОЭВМ

ОТЧЕТ

по лабораторной работе №6
по дисциплине «Статистические методы обработки экспериментальных данных»
Тема: Кластерный анализ. Метод k-средних.

Студент гр. 5381 _____ Лянгузов А. А.

Преподаватель _____ Середа В. И.

Санкт-Петербург
2019

Цель работы.

Освоение основных понятий и некоторых методов кластерного анализа.

Задание

Дано конечное множество из объектов, представленных двумя признаками (в качестве этого множества принимаем исходную двумерную выборку, сформированную ранее в лабораторной работе №4). Выполнить разбиение исходного множества объектов на конечное число подмножеств (кластеров) с использованием метода k-средних. Полученные результаты содержательно проинтерпретировать.

Основные теоретические положения.

Кластерный анализ (англ. *cluster analysis)* – это метод классификации многомерных наблюдений на основе определения сходства или близости (расстояния) между объектами. Цель кластерного анализа заключается в определении однородных в некотором смысле групп, которые называются кластерами.

Алгоритм кластерного анализа включает пять этапов.

- 1 этап. Представление исходных данных в виде матрицы (таблицы "объект – признак").
- 2 этап. Определение сходства объектов.
- 3 этап. Выбор метода объединения объектов в кластеры.
- 4 этап. Определение оптимального числа кластеров.
- 5 этап. Интерпретация кластеров и качества разбиения.

Меры расстояний:

Для того, чтобы сравнивать два объекта, необходимо иметь *критерий*, на основании которого будет происходить сравнение. Как правило, таким критерием является *расстояние* между объектами.

Есть множество мер расстояния, рассмотрим несколько из них:

Евклидово расстояние — наиболее распространенное расстояние. Оно является геометрическим расстоянием в многомерном пространстве:

$d(x,\ y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$Квадрат евклидова расстояния. Иногда может возникнуть желание возвести в квадрат стандартное евклидово расстояние, чтобы придать большие веса более отдаленным друг от друга объектам.

Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является просто средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако отметим, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

Расстояние Чебышева. Это расстояние может оказаться полезным, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением).

Алгоритм k-means (k-средних)

Наиболее простой, но в то же время достаточно неточный метод кластеризации в классической реализации. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k. Действие алгоритма таково, что он стремится минимизировать среднеквадратичное отклонение на точках каждого кластера. Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.

Экспериментальные результаты.

Двумерная выборка: n = 107

Таблица 1

v	501.00	369.00	344.00	473.00	426.00	528.00	497.00	467.00	506.00	431.00	454.00
E	130.40	84.30	86.80	137.90	121.10	163.40	147.30	140.50	158.40	125.00	131.10
v	371.00	482.00	393.00	441.00	463.00	440.00	481.00	340.00	468.00	397.00	496.00
E	89.20	139.90	103.20	122.80	129.10	128.50	135.20	85.10	142.00	108.60	143.10
v	434.00	541.00	352.00	438.00	453.00	423.00	351.00	525.00	409.00	469.00	386.00
E	122.30	146.80	87.70	134.90	119.50	131.10	89.00	165.90	121.00	131.50	95.50
v	505.00	436.00	488.00	449.00	493.00	512.00	472.00	423.00	465.00	351.00	359.00
E	137.50	114.30	134.10	124.50	129.70	169.90	134.20	130.80	140.70	102.90	71.90
v	457.00	467.00	400.00	418.00	492.00	434.00	510.00	392.00	463.00	459.00	397.00
E	126.40	135.10	114.60	118.60	137.50	110.50	140.60	82.70	125.00	145.40	106.80
v	424.00	436.00	429.00	398.00	493.00	522.00	518.00	463.00	437.00	386.00	493.00
E	119.00	116.70	112.90	109.00	154.50	154.50	144.40	121.20	121.80	105.80	151.20
v	414.00	480.00	585.00	562.00	508.00	421.00	463.00	422.00	406.00	544.00	345.00
E	113.50	153.90	177.70	175.90	159.00	117.80	136.70	122.90	110.10	166.70	95.90
v	478.00	393.00	437.00	448.00	458.00	422.00	468.00	430.00	371.00	543.00	471.00
E	126.60	122.80	115.10	121.90	121.70	115.70	144.90	104.30	91.90	155.40	143.90
v	475.00	521.00	353.00	437.00	362.00	490.00	484.00	459.00	480.00	482.00	522.00
E	132.00	139.60	98.00	118.40	111.70	139.90	140.40	136.70	153.30	148.20	143.80
v	576.00	390.00	514.00	442.00	421.00	443.00	438.00	429.00			
E	166.40	91.40	153.60	115.40	107.90	121.90	126.70	120.90			

Обработка результатов эксперимента.

1. Масштабирование выборки.

   Масштабирование выборки осуществляется по формулам:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \ \forall i$$

$$y_i = \frac{y_i - y_{min}}{y_{max} - y_{min}}, \ \forall i$$

v	0.6571 4286	0.1183 6735	0.0163 2653	0.5428 5714	0.3510 2041	0.7673 4694	0.6408 1633	0.5183 6735	0.6775 5102	0.3714 2857	0.4653 0612
E	0.5529 301	0.1172 023	0.1408 318	0.6238 185	0.4650 284	0.8648 393	0.7126 654	0.6483 932	0.8175 803	0.5018 904	0.5595 463
v	0.1265 3061	0.5795 9184	0.2163 2653	0.4122 4490	0.5020 4082	0.4081 6327	0.5755 1020	0.0000 0000	0.5224 4898	0.2326 5306	0.6367 3469
E	0.1635 161	0.6427 221	0.2958 412	0.4810 964	0.5406 427	0.5349 716	0.5982 987	0.1247 637	0.6625 709	0.3468 809	0.6729 679
v	0.3836 7347	0.8204 0816	0.0489 7959	0.4000 0000	0.4612 2449	0.3387 7551	0.0448 9796	0.7551 0204	0.2816 3265	0.5265 3061	0.1877 5510
E	0.4763 705	0.7079 395	0.1493 384	0.5954 631	0.4499 055	0.5595 463	0.1616 257	0.8884 688	0.4640 832	0.5633 270	0.2230 624
v	0.6734 6939	0.3918 3673	0.6040 8163	0.4448 9796	0.6244 8980	0.7020 4082	0.5387 7551	0.3387 7551	0.5102 0408	0.0448 9796	0.0775 5102
E	0.6200 378	0.4007 561	0.5879 017	0.4971 645	0.5463 138	0.9262 760	0.5888 469	0.5567 108	0.6502 836	0.2930 057	0.0000 000
v	0.4775 5102	0.5183 6735	0.2448 9796	0.3183 6735	0.6204 0816	0.3836 7347	0.6938 7755	0.2122 4490	0.5020 4082	0.4857 1429	0.2326 5306
E	0.5151 229	0.5973 535	0.4035 917	0.4413 989	0.6200 378	0.3648 393	0.6493 384	0.1020 794	0.5018 904	0.6947 070	0.3298 677
v	0.3428 5714	0.3918 3673	0.3632 6531	0.2367 3469	0.6244 8980	0.7428 5714	0.7265 3061	0.5020 4082	0.3959 1837	0.1877 5510	0.6244 8980
E	0.4451 796	0.4234 405	0.3875 236	0.3506 616	0.7807 183	0.7807 183	0.6852 552	0.4659 735	0.4716 446	0.3204 159	0.7495 274
v	0.3020 4082	0.5714 2857	1.0000 0000	0.9061 2245	0.6857 1429	0.3306 1224	0.5020 4082	0.3346 9388	0.2693 8776	0.8326 5306	0.0204 0816
E	0.3931 947	0.7750 473	1.0000 000	0.9829 868	0.8232 514	0.4338 374	0.6124 764	0.4820 416	0.3610 586	0.8960 302	0.2268 431
v	0.5632 6531	0.2163 2653	0.3959 1837	0.4408 1633	0.4816 3265	0.3346 9388	0.5224 4898	0.3673 4694	0.1265 3061	0.8285 7143	0.5346 9388
E	0.5170 132	0.4810 964	0.4083 176	0.4725 898	0.4706 994	0.4139 887	0.6899 811	0.3062 382	0.1890 359	0.7892 250	0.6805 293
v	0.5510 2041	0.7387 7551	0.0530 6122	0.3959 1837	0.0897 9592	0.6122 4490	0.5877 5510	0.4857 1429	0.5714 2857	0.5795 9184	0.7428 5714

E	0.5680529	0.6398866	0.2466919	0.4395085	0.3761815	0.6427221	0.6474480	0.6124764	0.7693762	0.7211720	0.6795841
v	0.96326531	0.20408163	0.71020408	0.41632653	0.33061224	0.42040816	0.40000000	0.36326531			
E	0.8931947	0.1843100	0.7722117	0.4111531	0.3402647	0.4725898	0.5179584	0.4631380			

2. K-means кластеризация.

   Количество кластеров взяли равным 7.

   В качестве начальных центров кластеров взяли рандомные пары (v, E).

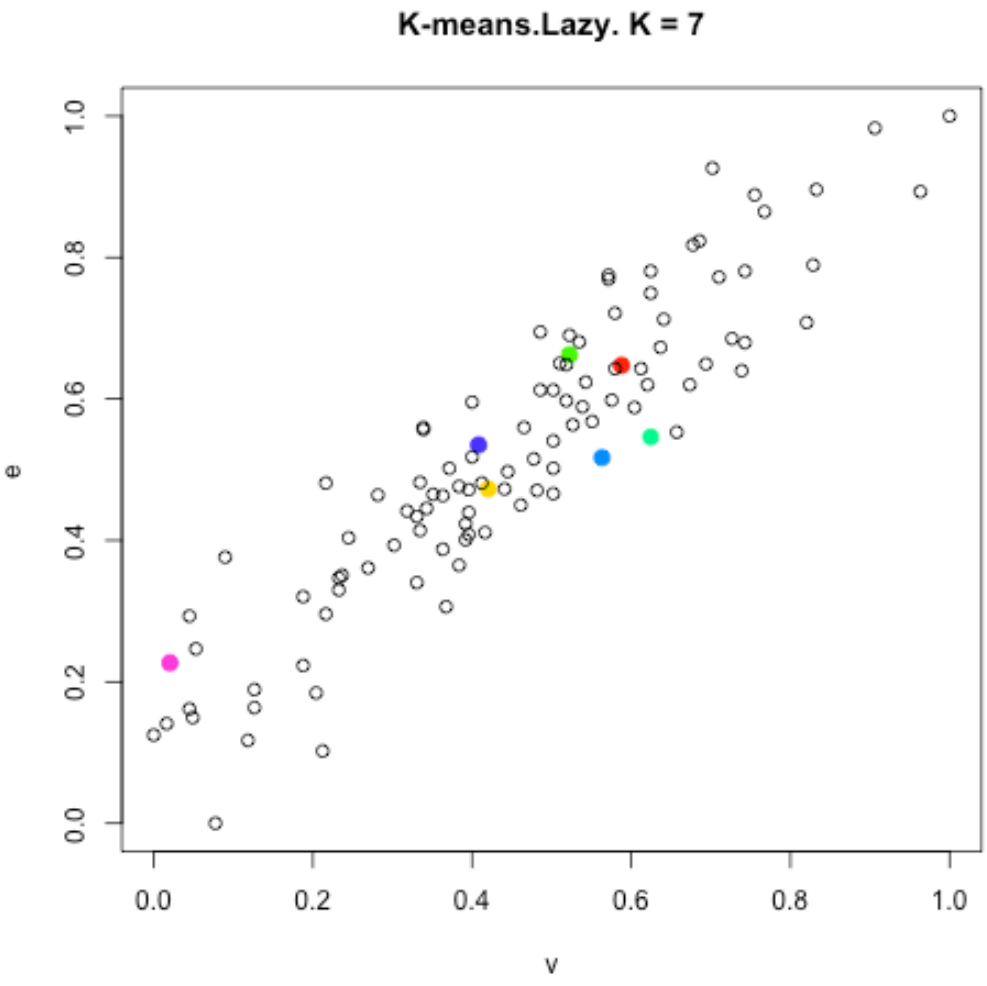
   Пересчет центров кластеров после распределения всех пар осуществляется по формулам:

$$\bar{x} = \frac{\sum_{i=1}^{n_k} x_i}{n_k}$$

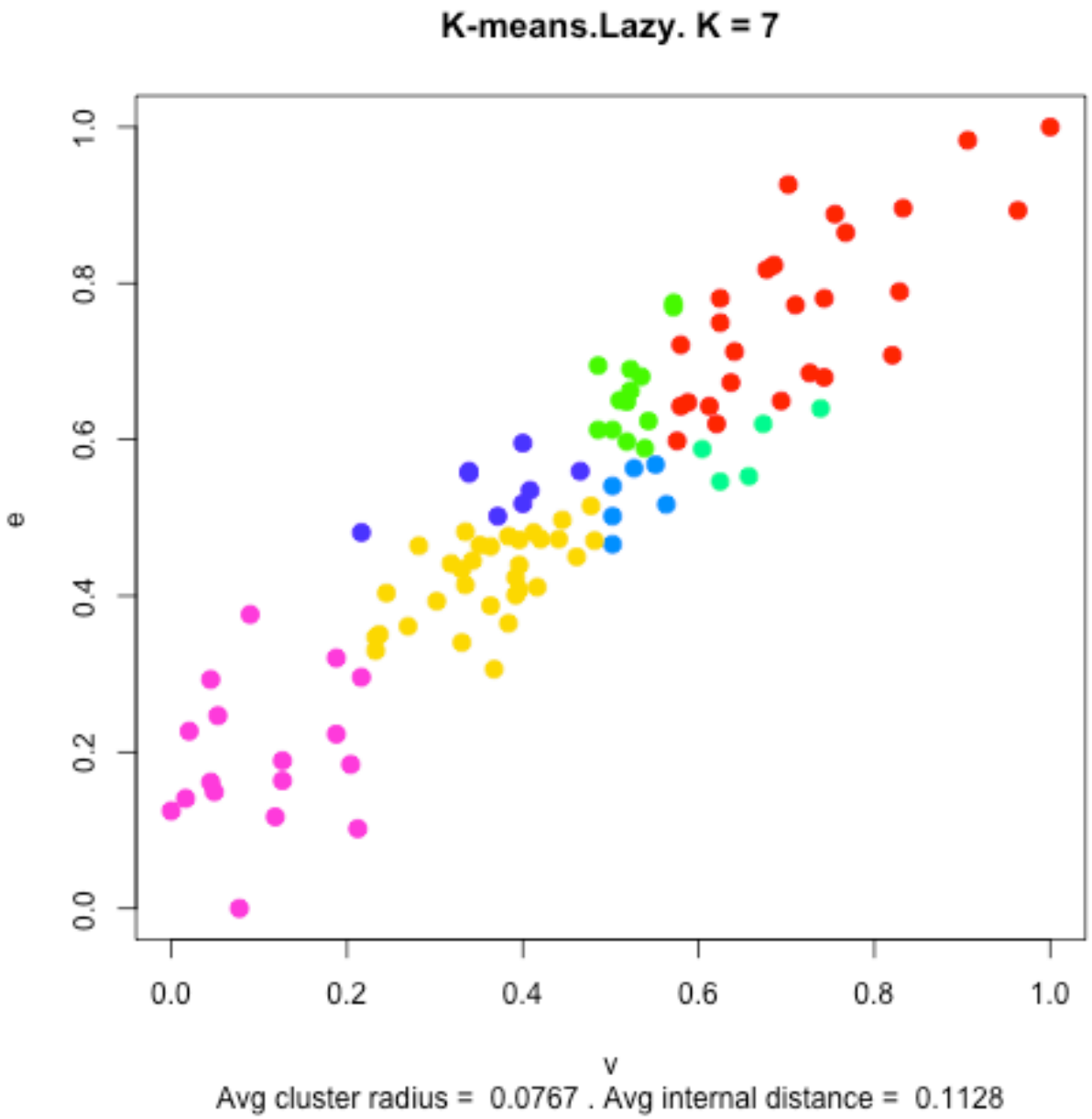
$$\bar{y} = \frac{\sum_{i=1}^{n_k} y_i}{n_k}, \text{ где } n_k - \text{количество точек в кластере}.$$

   Результаты представим графически.

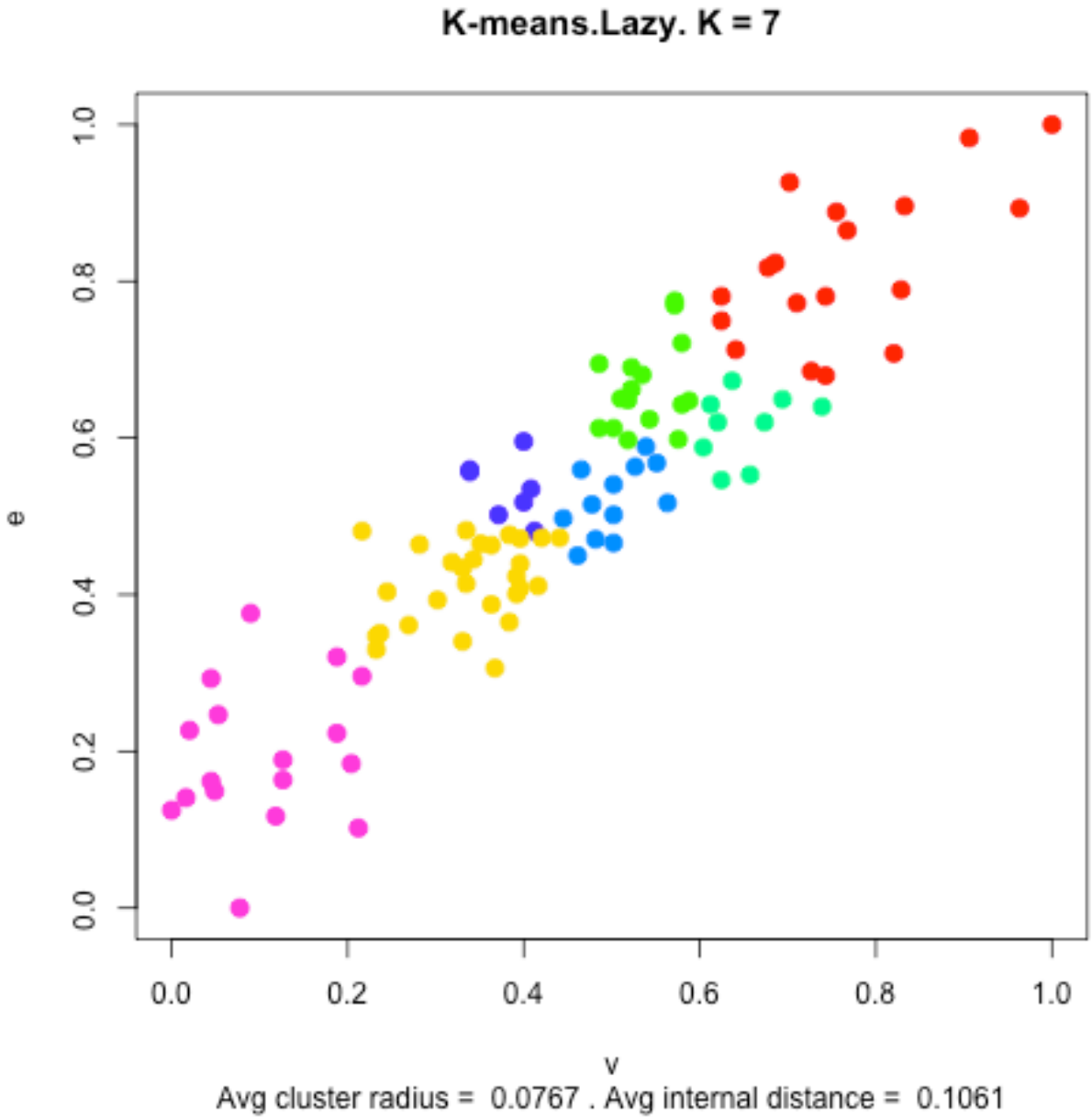
   1) Случайные центры:



2) Кластеризация по данным центрам:

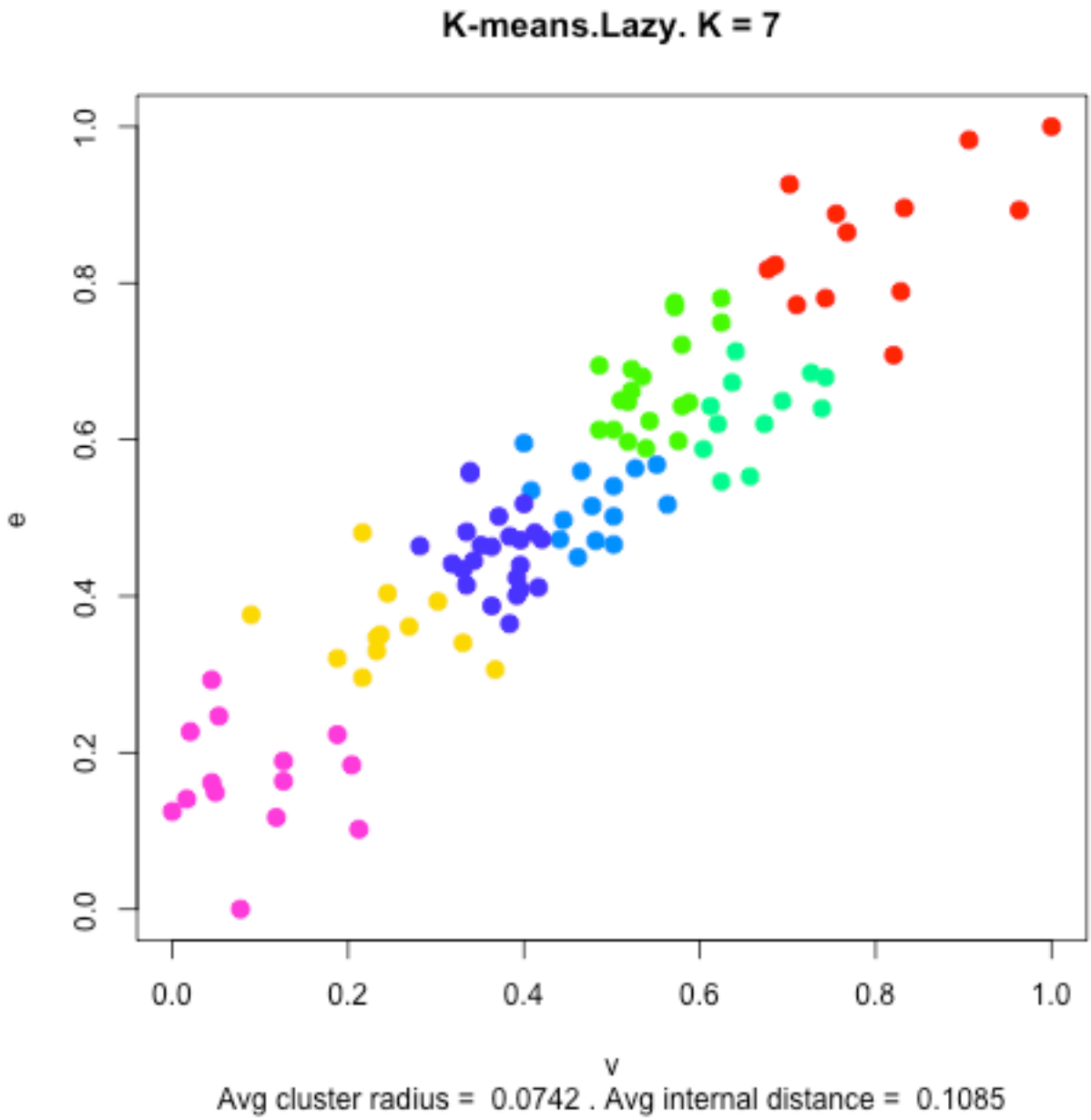


3) Пересчет центров и кластеризация:



4) Пересчет центров осуществляется пока центры не перестанут меняться.

В данном случае конечный вариант:



Вывод.

В ходе выполнения лабораторной работы ознакомились кластерным анализом, в частности, k-means - методом *k*-средних.

К недостаткам k-means можно отнести:

- необходимость заранее знать количество кластеров;
- чувствительность к выбору начальных центров кластеров.

В отчете использовался вариант k-means, когда пересчет центров кластеров осуществляется после распределения всех пар (v, E) по кластерам, пока центры не перестанут меняться (кодовое название: “Lazy”). Однако, разработанный скрипт также содержит другой вариант метода k-средних, где пересчет центров происходит после каждого добавления точки (кодовое название: “Long”).

В результате работы алгоритма k-средних, выборка была разделена на 7 кластеров.