МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра математического обеспечения и применения ЭВМ

ОТЧЕТ

по лабораторной работе №6

по дисциплине «Статистические методы обработки экспериментальных данных»

Тема: Кластерный анализ. Метод k-средних.

Студентка гр. 8382 _____ Звегинцева Е.Н.

Студент гр. 8382 _____ Мирончик П.Д.

Преподаватель _____ Середа А.-В.И.

Санкт-Петербург
2022

Цель работы.

Освоение основных понятий и некоторых методов кластерного анализа, в частности, метода k-means.

Основные теоретические положения.

Кластерный анализ – многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы.

К характеристикам кластера относятся в частности: центр, радиус; среднеквадратическое отклонение; размер кластера.

Центр кластера – это среднее геометрическое место точек, принадлежащих кластеру, в пространстве данных.

Радиус кластера – максимальное расстояние точек, принадлежащих кластеру, от центра кластера.

Нормировка, т.е. стандартизация, переменных применяется для того, чтобы характеристики имели один масштаб, в следствии чего было возможно корректное разбиение на кластеры. В данной работе для нормировки была использована формула

$$z = \frac{x - \bar{x}}{\sigma_x}$$

Где σ_x – стандартное отклонение переменной, а $\bar{x}$ – ее среднее значение.

Евклидово расстояние (способ определения расстояния между наблюдениями):

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Алгоритм k-means – это наиболее популярный метод кластеризации, который разделяет определенный набор данных на заданное пользователем число кластеров k. В начале классификации задается число k классов и выбира-

ются k точек, которые будут служить центрами кластеров. Для каждого наблюдения из исходной выборки вычисляются расстояния до центров кластеров. Наблюдение распределяется в кластер, центр которого находится ближе всего к нему.

Возможны две разновидности метода *k* -means. Первая предполагает пересчет центра кластера после каждого изменения его состава, а вторая —лишь после завершения цикла.

Наилучшим разбиением считается такое, при котором достигается экстремальное (минимальное или максимальное) значение выбранного функционала качества.

В качестве таких функционалов могут быть использованы:

- Сумма по всем кластерам квадратов расстояний элементов кластеров до центров соответствующих кластеров:

$$F_1 = \sum_k \sum_{i=1}^{N_k} ||x_{k,i} - c_k||^2 = \sum_k \sum_{i=1}^{N_k} \left((\nu_{k,i} - \nu_{c_k})^2 + (E_{k,i} - E_{c_k})^2 \right)$$

- Сумма по всем кластерам внутрикластерных расстояний между элементами кластеров:

$$F_2 = \sum_k \sum_{i=1}^{N_k} \sum_{j=i+1}^{N_k} ||x_{k,i} - x_{k,j}||^2 = \sum_k \sum_{i=1}^{N_k} \sum_{j=i+1}^{N_k} \left((\nu_{k,i} - \nu_{k,j})^2 + (E_{k,i} - E_{k,j})^2 \right)$$

- Сумма по всем кластерам внутрикластерных дисперсий (относительно центров кластеров):

$$F_3 = \sum_k \sum_{i=1}^{N_k} \sigma_{k,i}^2 = \sum_k \frac{1}{N_k} \sum_{i=1}^{N_k} (x_{k,i} - c_k)^2 = \sum_k \frac{1}{N_k} \sum_{i=1}^{N_k} \left((\nu_{k,i} - \nu_{c_k})^2 + (E_{k,i} - E_{c_k})^2 \right)$$

Оптимальным следует считать разбиение, при котором сумма внутрикластерных (внутригрупповых) дисперсий будет минимальной.

Постановка задачи.

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.

Выполнение работы.

Для выполнения данной работы была использована выборка, сформированная в первой лабораторной работе.

Нам нужно корректно реализовать методы кластерного анализа, для чего мы нормализуем нашу выборку по формуле:

$$z = \frac{x - \bar{x}}{\sigma_x}$$

Где σ_x – стандартное отклонение переменной, а $\bar{x}$ – ее среднее значение.

Нормализированное множество представлено на рис.1 в таблице, а также на рис.2 на диаграмме рассеяния.

	0	1
0	-0.893663	-1.027792
1	1.325225	1.241278
2	-1.333615	-1.477776
3	-1.716182	-0.826735
4	-0.147657	0.049298
...	...	...
109	1.000043	0.781720
110	0.368808	0.714701
111	0.789631	1.222130
112	1.765177	1.806152
113	0.158396	-0.314519

114 rows × 2 columns

Рисунок 1 – таблица, сгенерированная программой

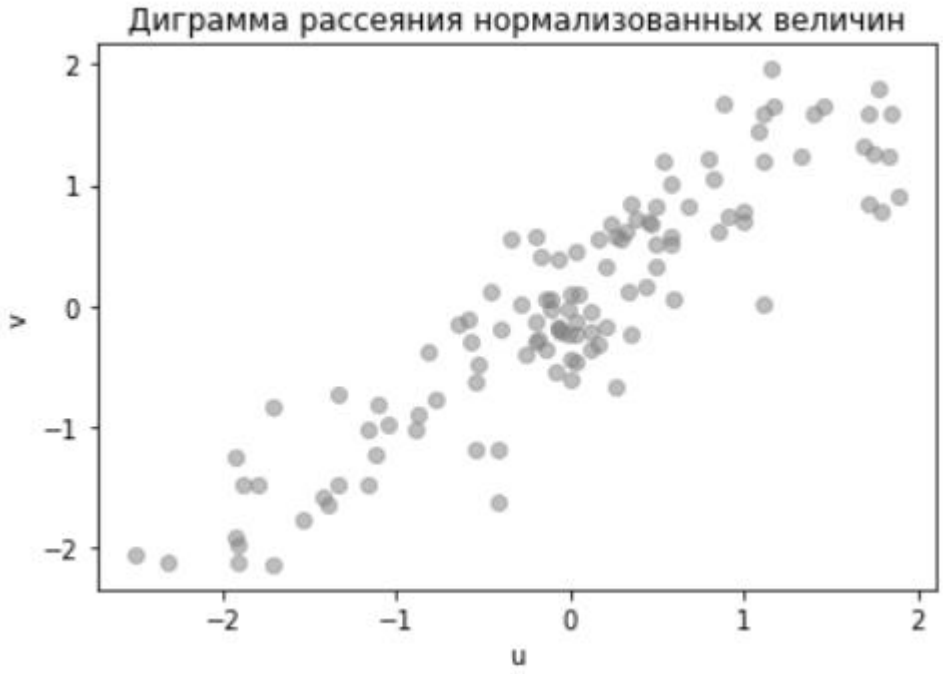


Рисунок 2 – диаграмма рассеяния нормализованных величин

Далее мы определяем верхнюю оценку количества кластеров:

$$\bar{k} = [\sqrt{N/2}]$$

$$\bar{k} = 7$$

Среди наблюдаемых значений выборки отбираются случайным образом начальные центры кластеров, в связи с тем, что проблематично определить визуально, где они точно находятся. Далее эти центры были отмечены крестиками на рис.3.

	0	1
0	0.043626	0.101956
1	-1.716182	-0.826735
2	1.688664	1.327445
3	-0.109401	-0.027295
4	-0.166786	0.403541
5	0.330551	0.121104
6	1.172198	1.657753

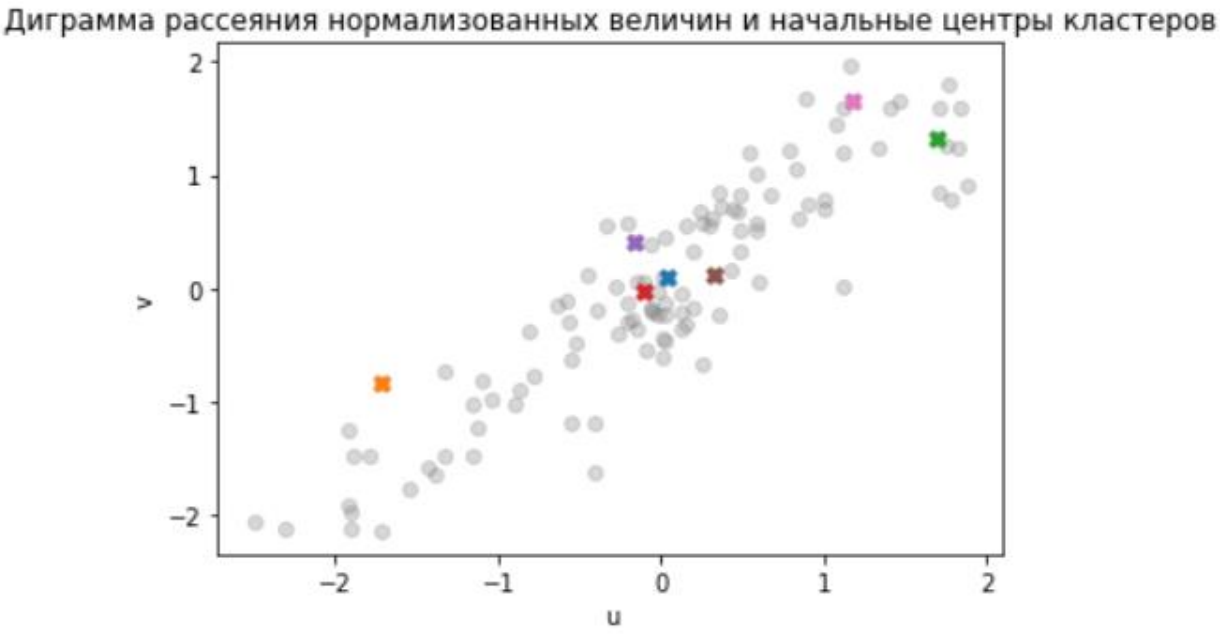


Рисунок 3 – Диаграмма рассеяния с начальными центрами классов

Далее были реализованы функции для осуществления метода кластеризации *k-means:*

- Функция, определяющая индекс кластера, котору принадлежит данное наблюдение (С помощью евклидова расстояния выбирается наименьшее расстояние).

$$clust(x) = argmin_k \sqrt{(v - v_{c_k})^2 + (E - E_{c_k})^2}$$

- Функция, осуществляющая пересчет центров кластеров (среднее геометрическое место точек, принадлежащих кластеру). В случае пустоты кластера – центр не пересчитывается.

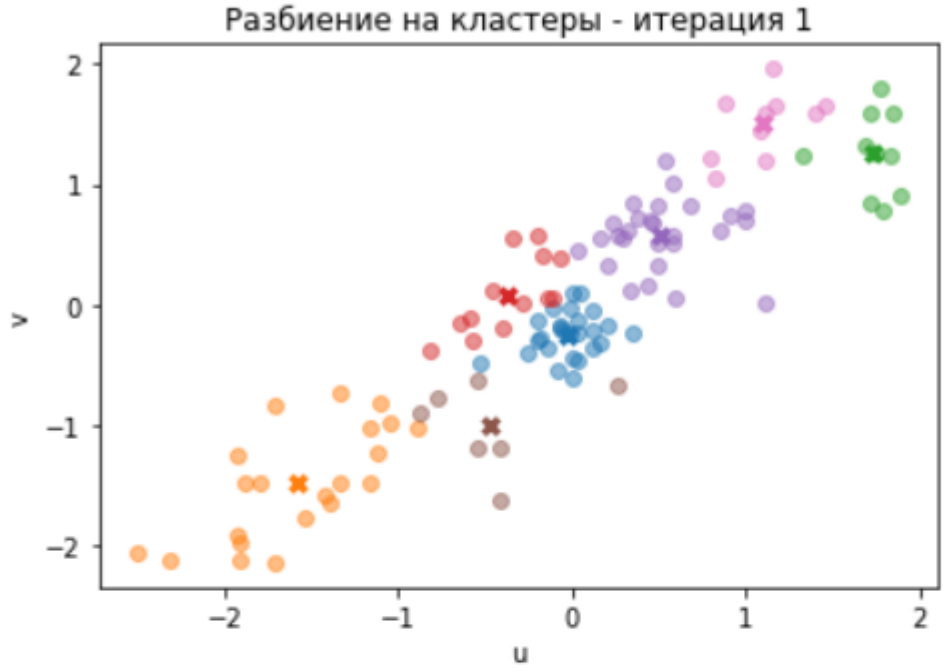
$$(\nu_{c_k}, E_{c_k}) = \left(\frac{1}{N_k} \sum_{i=1}^{N_k} \nu_{k,i}, \frac{1}{N_k} \sum_{i=1}^{N_k} E_{k,i} \right)$$

- Функция-перебор элементов для распределения в кластеры
- Три функции для функционалов качества полученного разбиения. Т.е. функции для сумм по всем кластерам квадратов расстояний элементов кластеров до центров соответствующих кластеров, внутрикластерных расстояний между элементами кластеров и внутрикластерных дисперсий.

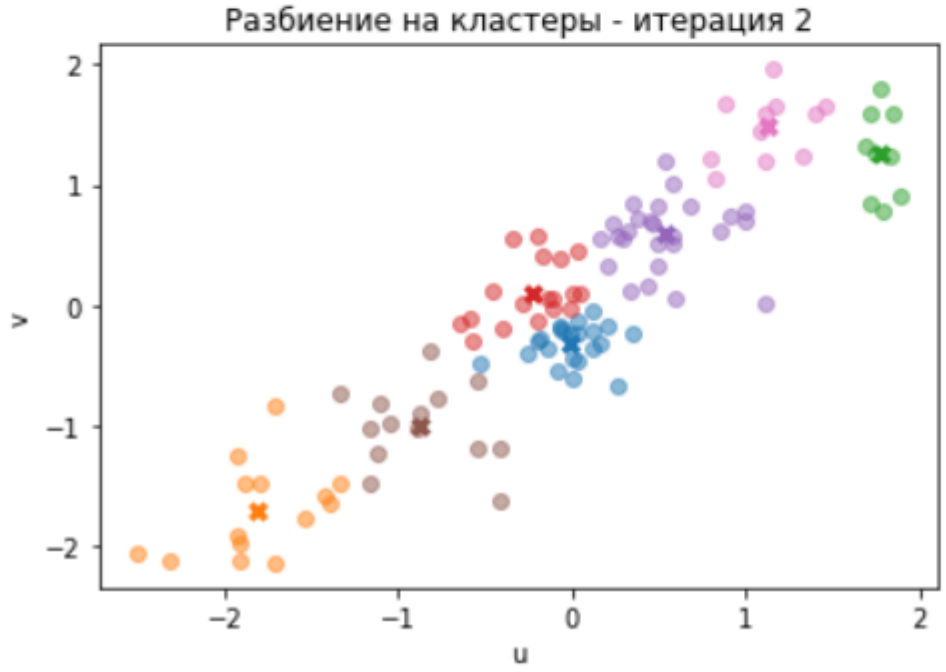
Диаграмма рассеяния выводится на каждом шаге, с пометками центров кластеров, вместе со значениями функционалов качества. При совпадении результата с предыдущей итерацией – алгоритм завершает работу.

Работа алгоритма с пересчетом центров кластеров после каждого изменения состава кластеров:

```
Итерация  1
F1: 19.393306676626427, F2: 535.0259554066845, F3: 1.1940458236227163
```

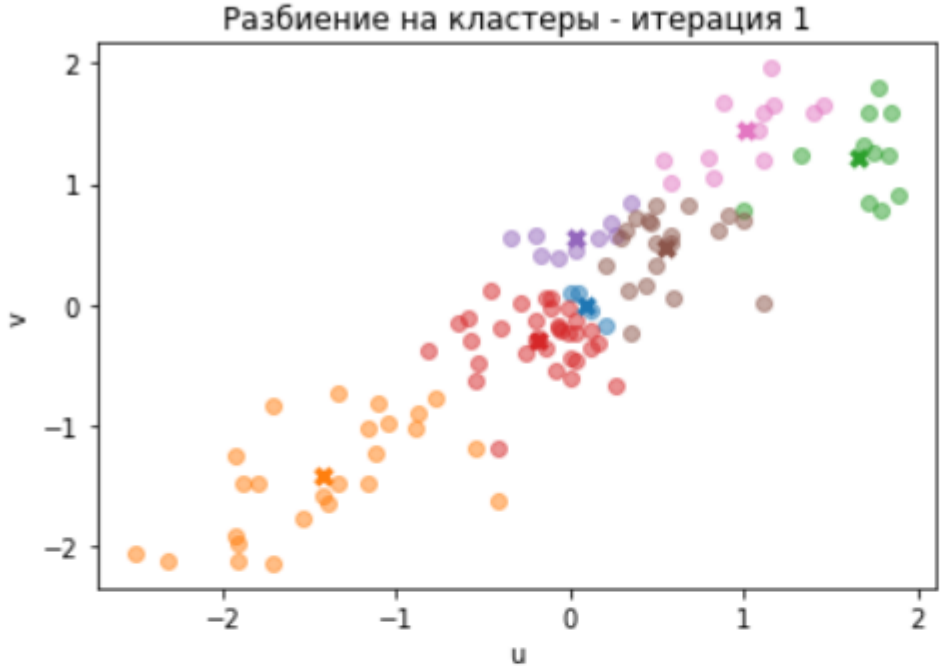



```
Итерация  2
F1: 15.224404173149178, F2: 446.29854197798693, F3: 0.9675215872510854
```

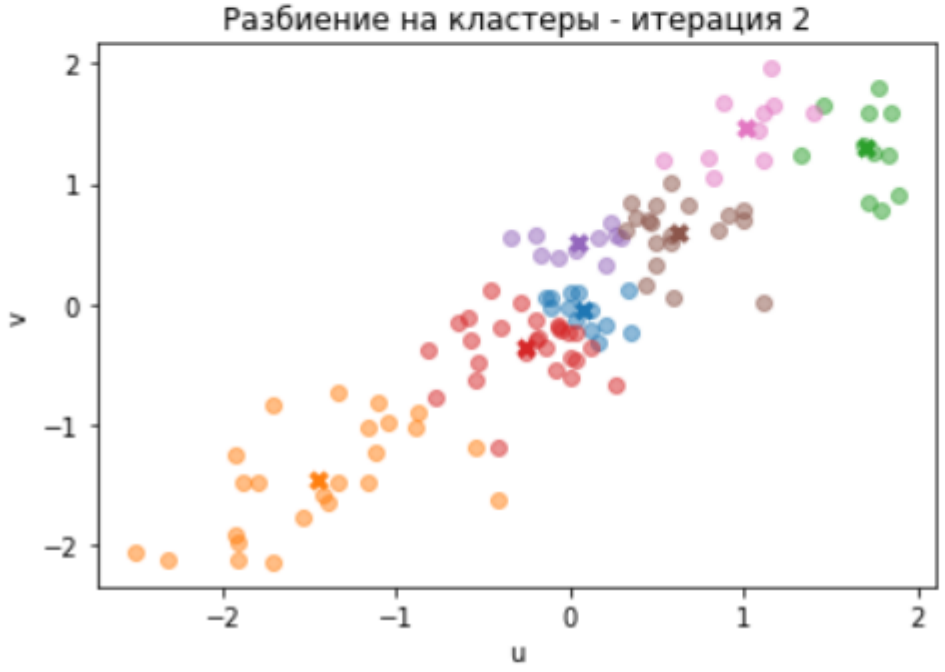


Работа алгоритма с пересчетом центров кластеров после просмотра всех выборочных значений:

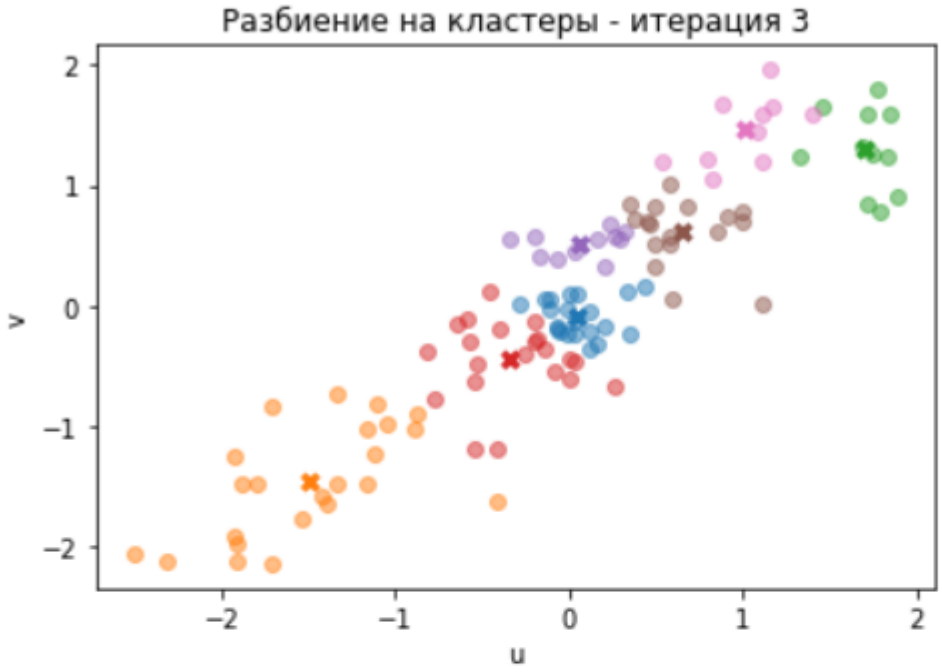
```
Итерация  1
F1: 23.7033305084117, F2: 675.2771600395128, F3: 1.1761297669007267
```



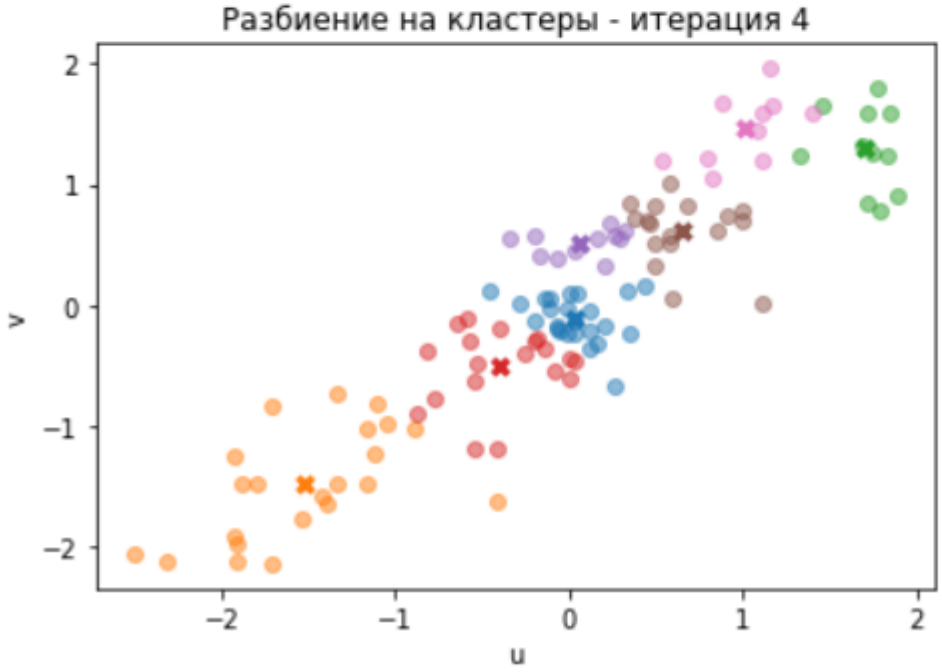

```
Итерация  2
F1: 21.232130254221598, F2: 571.1797166921232, F3: 1.0908119936491505
```



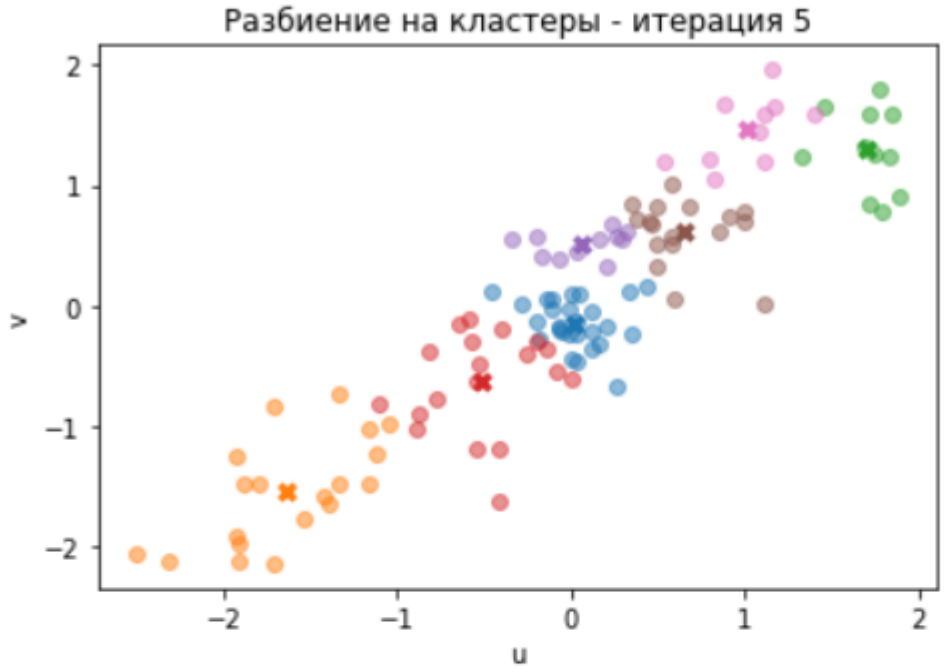
```
Итерация  3
F1: 20.40281719182281, F2: 517.4760858619246, F3: 1.110276713699098
```



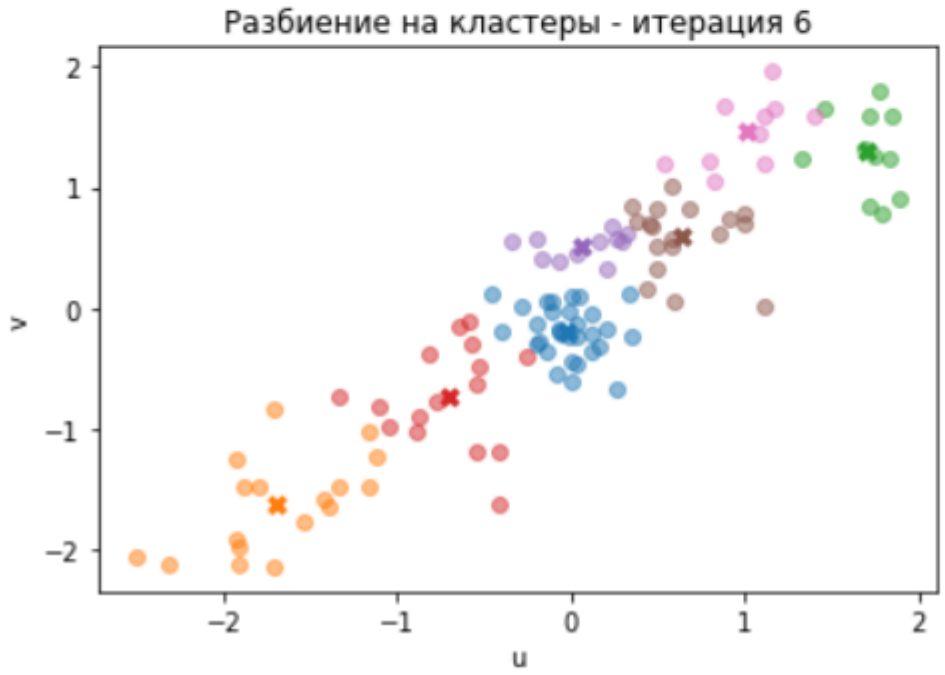

```
Итерация  4
F1: 19.9064949555773, F2: 507.7614462416363, F3: 1.1137733710851783
```



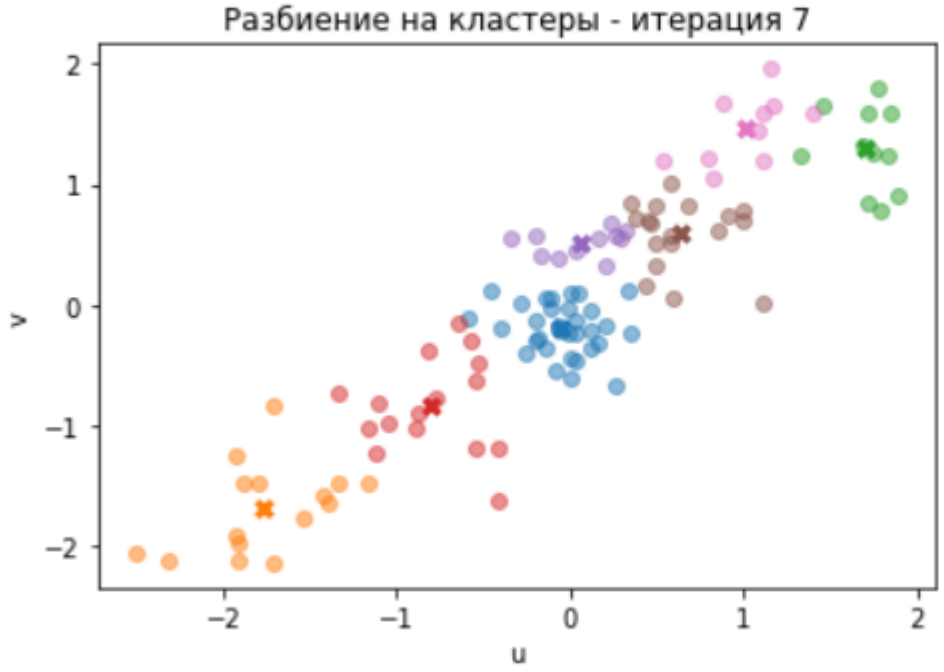
```
Итерация  5
F1: 18.72063427958748, F2: 489.9955296177793, F3: 1.1061093851714259
```



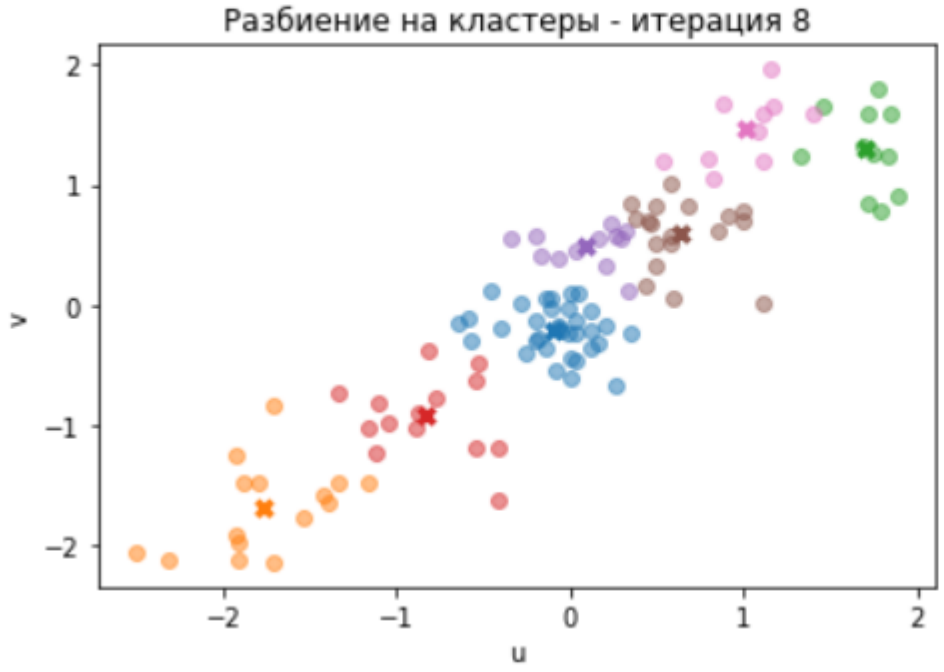

```
Итерация  6
F1: 16.93317071284603, F2: 475.6917448128395, F3: 1.0583422339065112
```



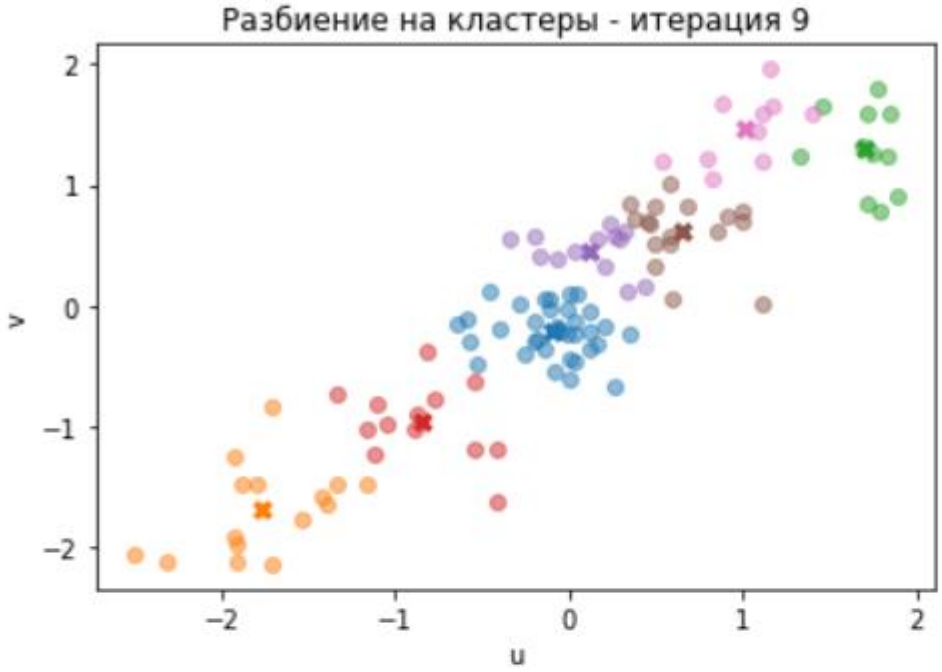
Итерация 7
F1: 15.770898625110433, F2: 476.7659677939061, F3: 1.0021721685887917



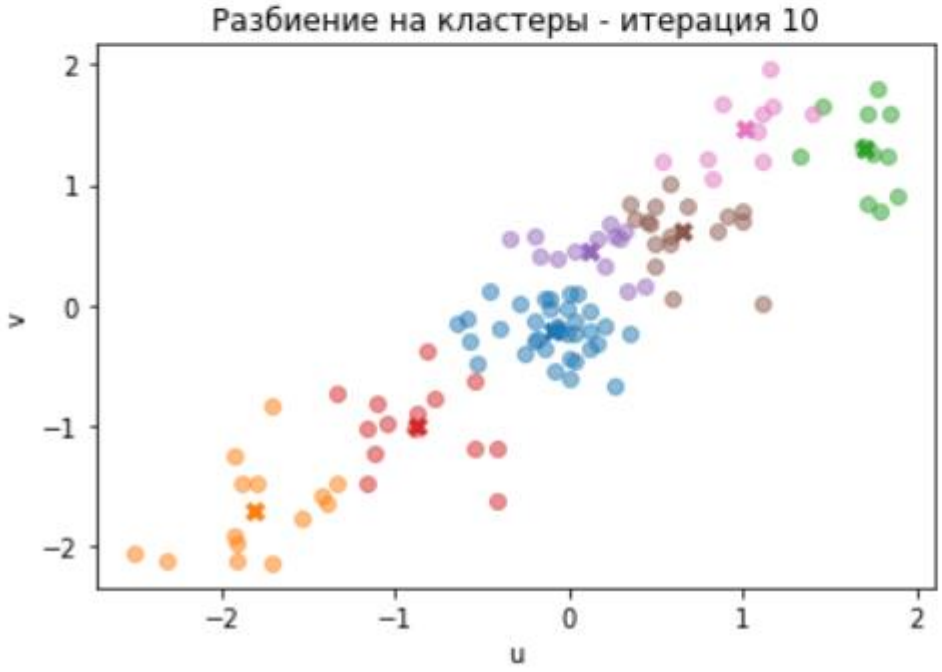
Итерация 8
F1: 15.344529151175895, F2: 480.27571379652784, F3: 0.9848180648039133



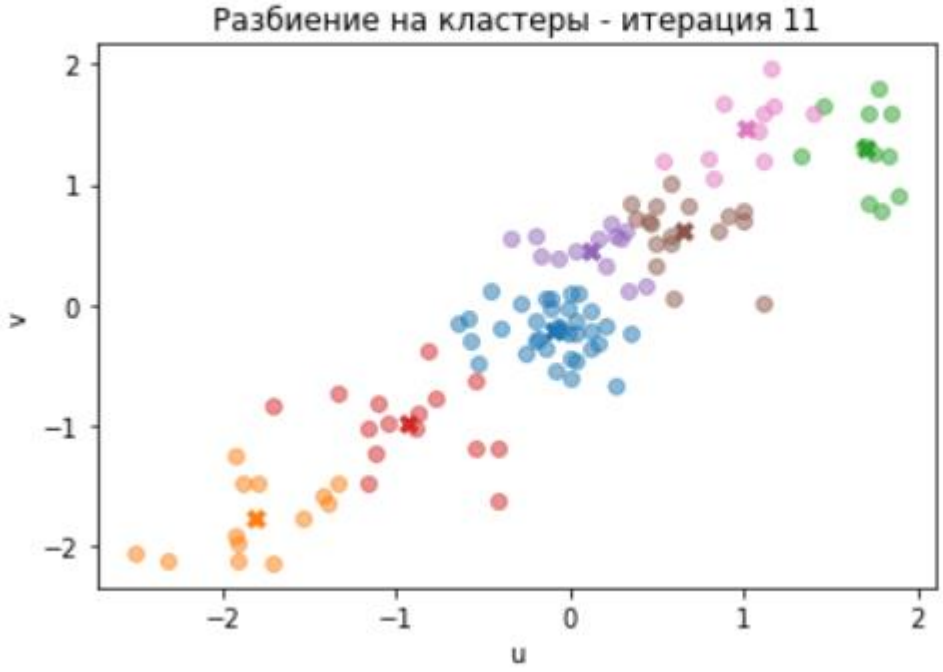
```
Итерация  9
F1: 15.256405642454517, F2: 487.4228601423251, F3: 0.9833061131093985
```




```
Итерация  10
F1: 15.155278326180667, F2: 485.79302969565157, F3: 0.9816883548460924
```



```
Итерация  11
F1: 15.01298445250833, F2: 485.5962468662807, F3: 0.9696943163285154
```



Как мы можем заметить, алгоритм, пересчитывающий центры после каждого изменения состава кластеров, работает гораздо быстрее, но результат полученный двумя алгоритмами различается. В первом случае мы получили немного большее расстояние от элементов до центров кластеров (функционал 1), но меньшее внутрикластерное расстояние и внутрикластерные дисперсии, относительно центров (функционал 2 и 3).

Выводы.

В ходе выполнения лабораторной работы, были освоены основные понятия кластерного анализа, в частности, метода *k-means,* с помощью которого было осуществлено распределение наблюдений по семи кластерам. Были реализованы две вариации алгоритма. Алгоритм с пересчетом центров кластеров после каждого изменения их состава - сходится гораздо быстрее (2 итерации) и захватывает значения имеющие меньшее внутрикластерное расстояние и внутрикластерные дисперсии, относительно центров. Алгоритм с пересчетом центров кластеров после просмотра всех элементов выборки захватывает значения менее удаленные от центра выборки.