

ПИиКТ

Современные инструменты анализа данных

Лабораторная работа №1

Выполнили: Нуруллаев Даниил Романович и

Балтабаев Дамир Темиржанович

Группы:P3214 и P3210

Преподаватель: Кустова Е. С.

Генерируем свои данные

import pandas as pd //для работы с анализом данных

import numpy as np //для использования математических функций

import matplotlib.pyplot as plt //для визуализации данных

from sklearn.cluster import DBSCAN

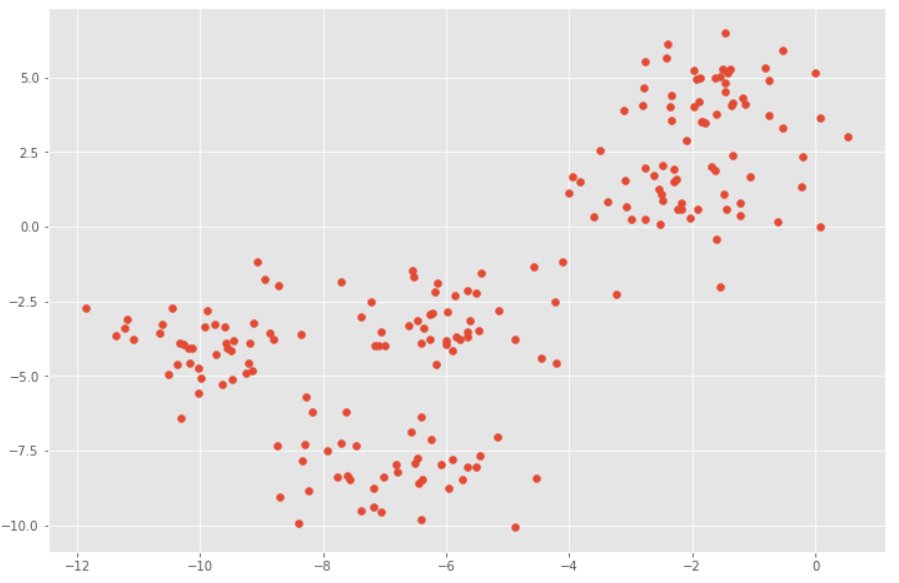
%matplotlib inline //показываем что нужно строить графики во все той же оболочке Jupyter

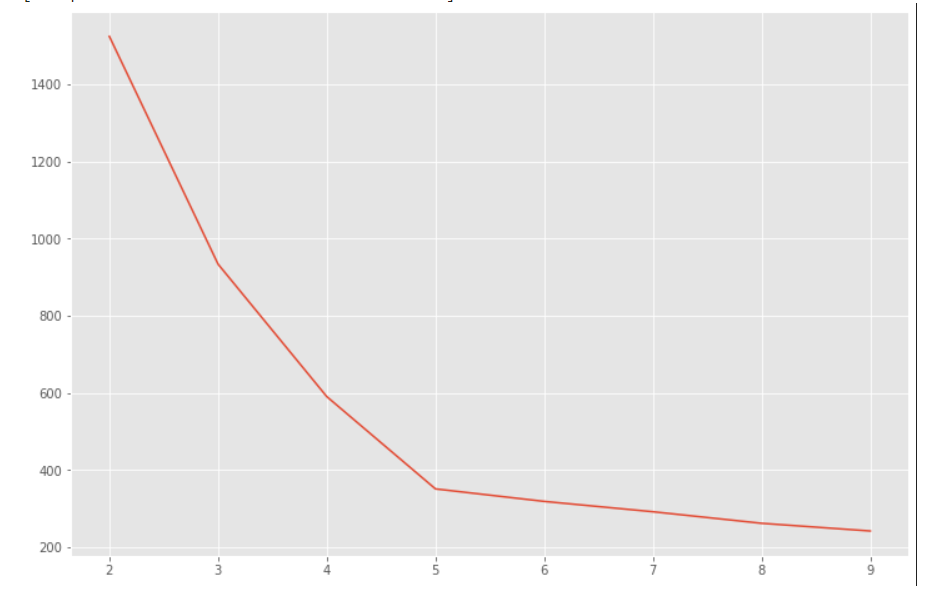
plt.style.use('ggplot') //для создания визуализации

plt.rcParams['figure.figsize']=(12,8)

X,y = make\_blobs(n\_samples=200, random\_state=1, centers=5) //создаем игрушечные данные, указываем кол-во 200, рандом стэйт любой для повторения результатов и 5 кластеров

plt.scatter(X[:,0],X[:,1]) //для создания диаграммы рассеивания





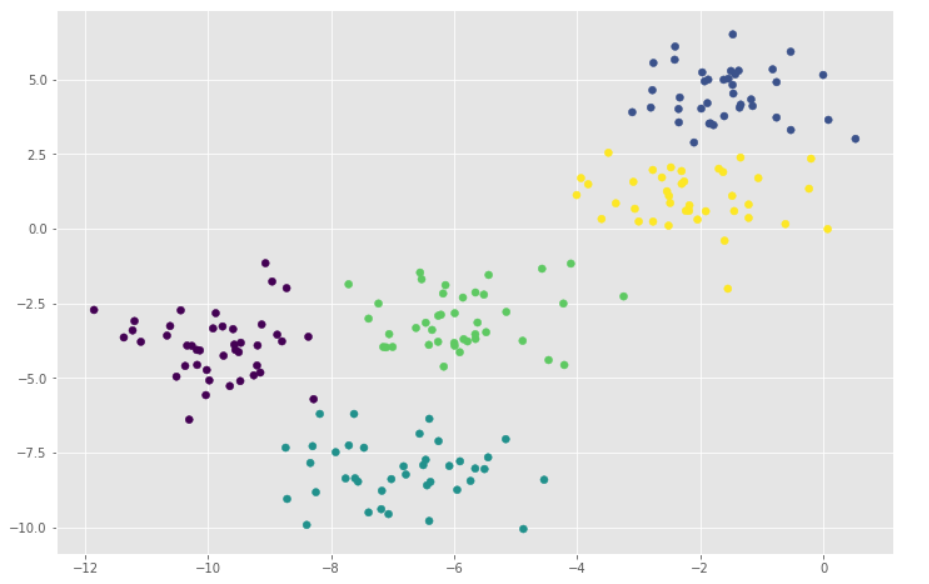
Используем метод KMean:

kmeansModel=KMeans(n\_clusters=5, random\_state=0)

kmeansModel.fit(X) //обучаем модель

labels = kmeansModel.labels\_ //получаем точки при помощи атрибута labels

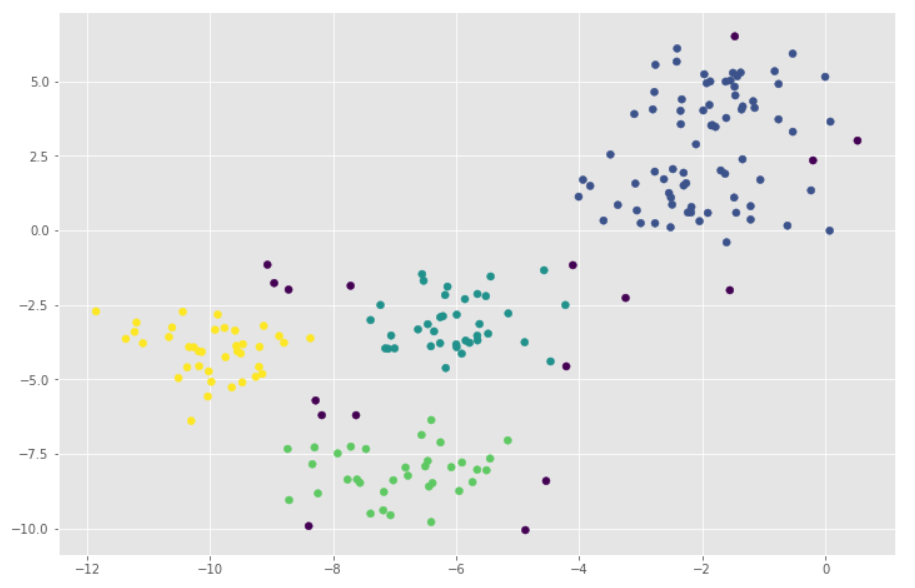
plt.scatter(X[:,0], X[:,1], c=labels) //для создания диаграммы рассеивания



Используем метод DBSCAN:

clustering = DBSCAN(eps=1, min\_samples=5).fit\_predict(X)

plt.scatter(X[:,0], X[:,1], c=clustering); //подбираем значения eps и min\_samples и используем DBSCAN



Теперь возьмем данные с сайта:

import pandas as pd

import numpy as np

from google.colab import drive

drive.mount('/content/drive')

data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/Mall\_Customers.csv', sep = ',')

frame = data.iloc[:,[3,4]]

criteries = []

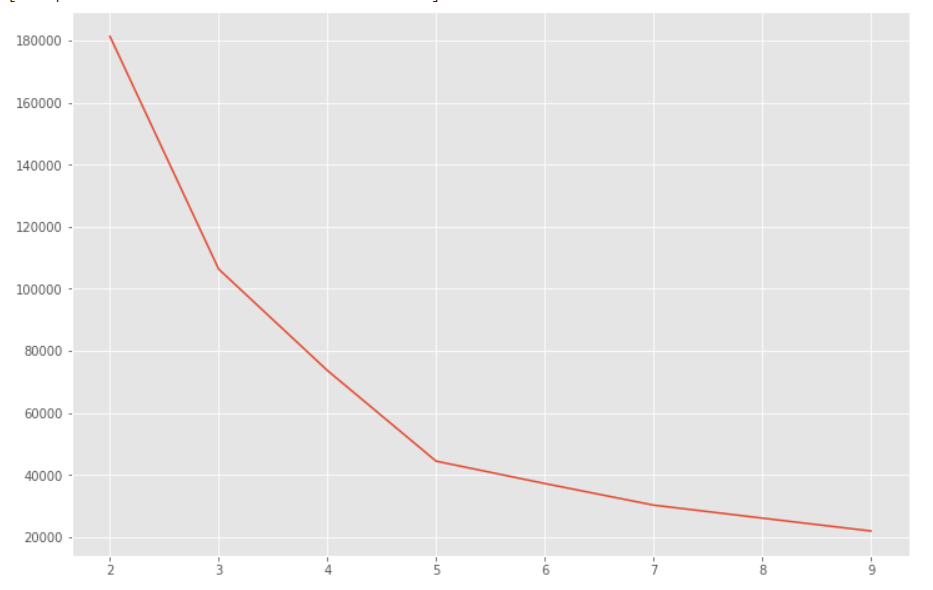
for k in range(2,10): //вычисляем значения kmeans для взятых точек

  kmeansModel=KMeans(n\_clusters=k, random\_state=3)

  kmeansModel.fit(frame) //обучаем модель

  criteries.append(kmeansModel.inertia\_) //добавляем вычисленный критерий, при помощи атрибута inertia

plt.plot(range(2,10), criteries)



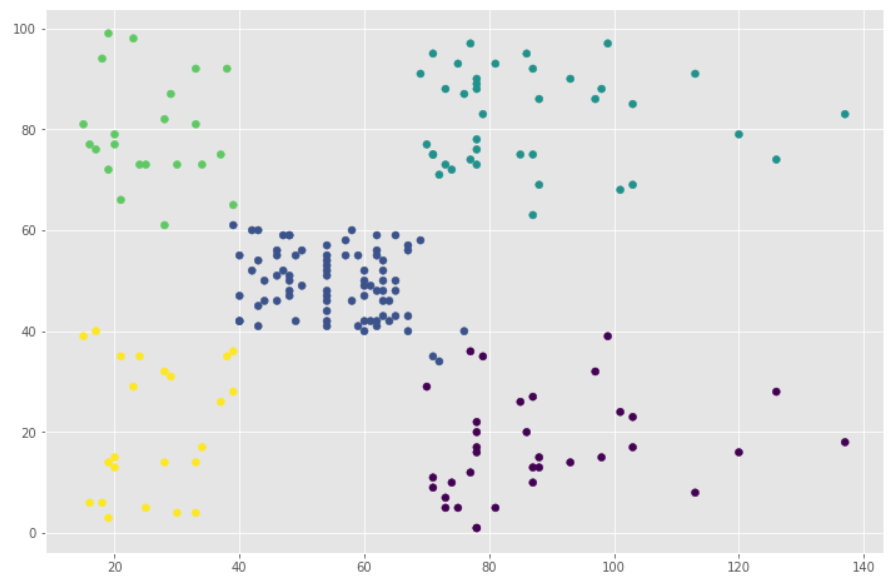
Используем метод KMean:

kmeansModel=KMeans(n\_clusters=5, random\_state=0) //применяем метод для 5 кластеров и смотрим результат

kmeansModel.fit(frame)

labels = kmeansModel.labels\_

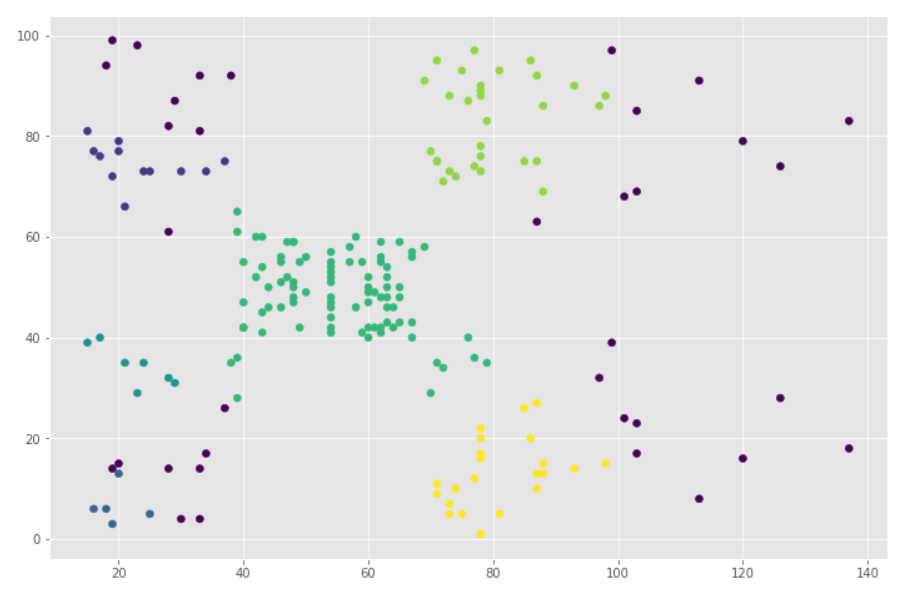
plt.scatter(frame.iloc[:,0], frame.iloc[:,1], c=labels)



Используем метод DBSCAN:

clustering = DBSCAN(eps=8, min\_samples=5).fit\_predict(frame)

plt.scatter(frame.iloc[:,0], frame.iloc[:,1], c=clustering);



Вывод: в данной лабораторной работе мы изучили метод кластеризации и выполнили с помощью нее работу на python.