Отчет по лабораторной работе №1

Задача №19

Яньшина А.Е.

В данной задаче необходимо провести кластерный анализ данных заработка людей. В рамках данной задачи проведено разделение людей на группы (кластеры) со схожей заработной платой.

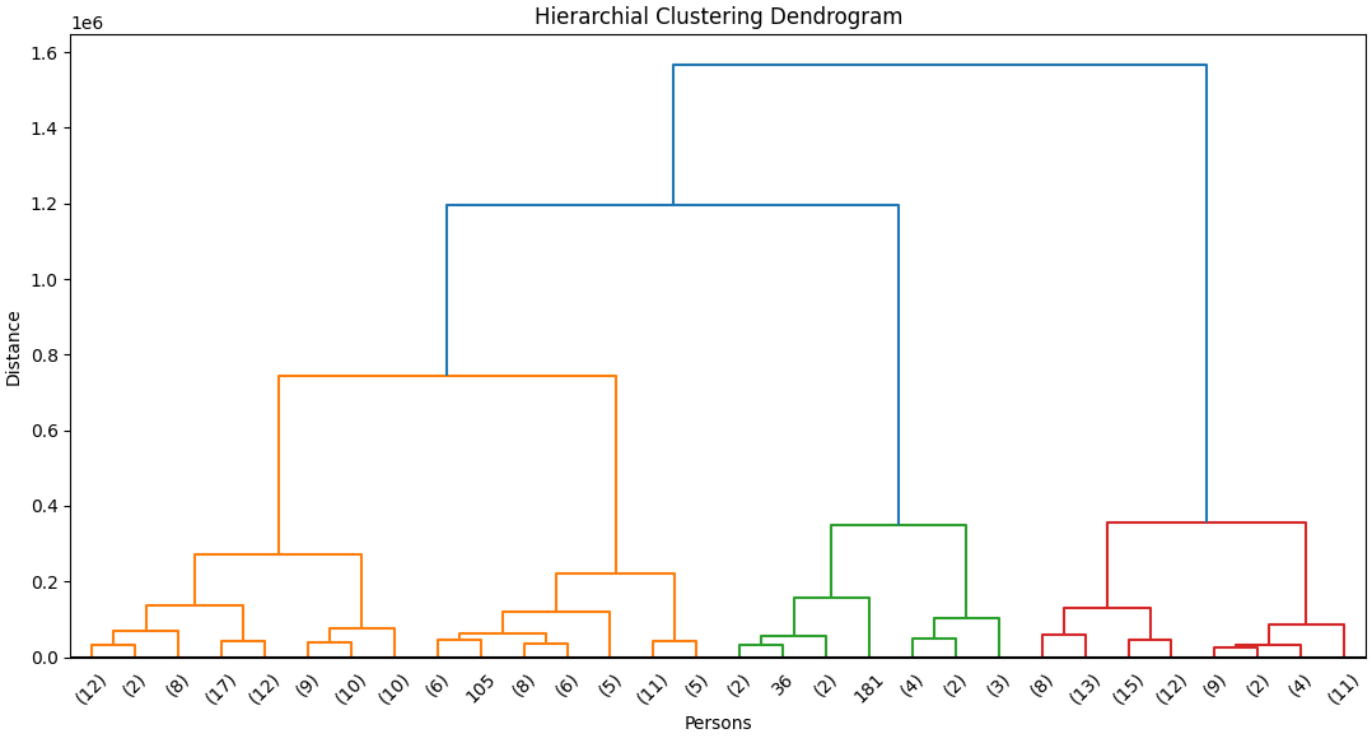
В качестве переменных, по которым производится кластерный анализ, используются age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week.

Так как в задании указана необходимость стандартизации данных, то первым делом стандартизируем полученные данные.

1) Иерархическая кластеризация

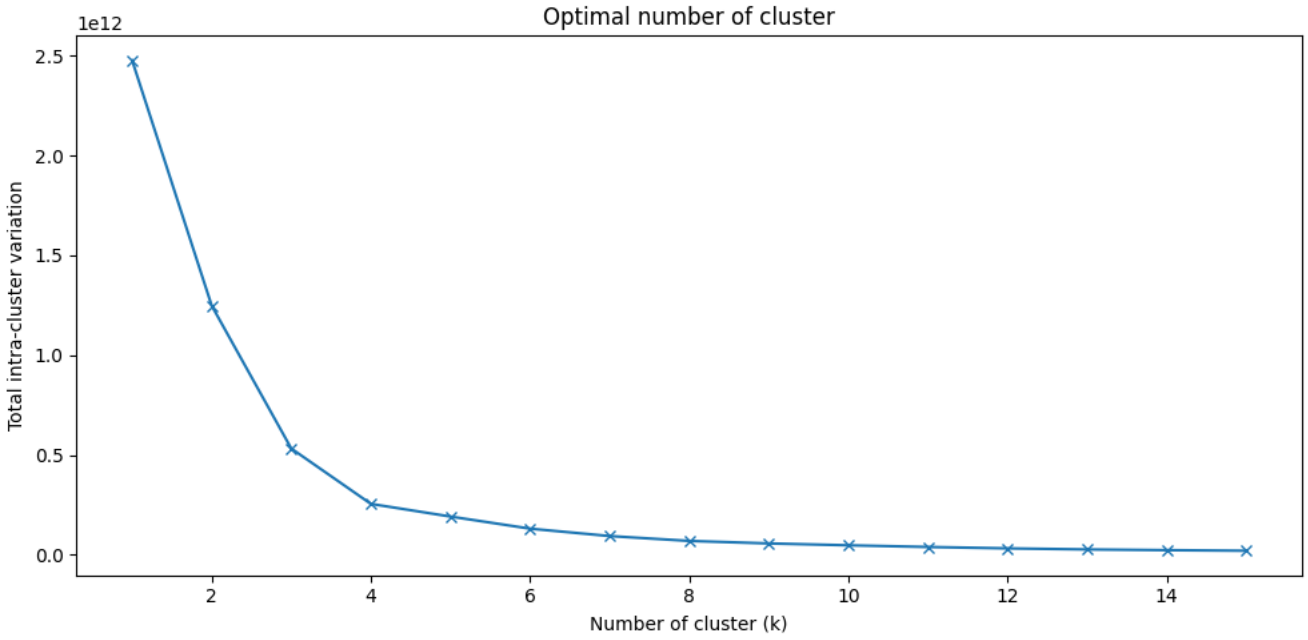
Для иерархической кластеризации используется пакет scipy.cluster.hierarchy.

При помощи которого строится дендрограмма.



Дистанция между объектами определяется при помощи метода Уорда (ось абцисс).

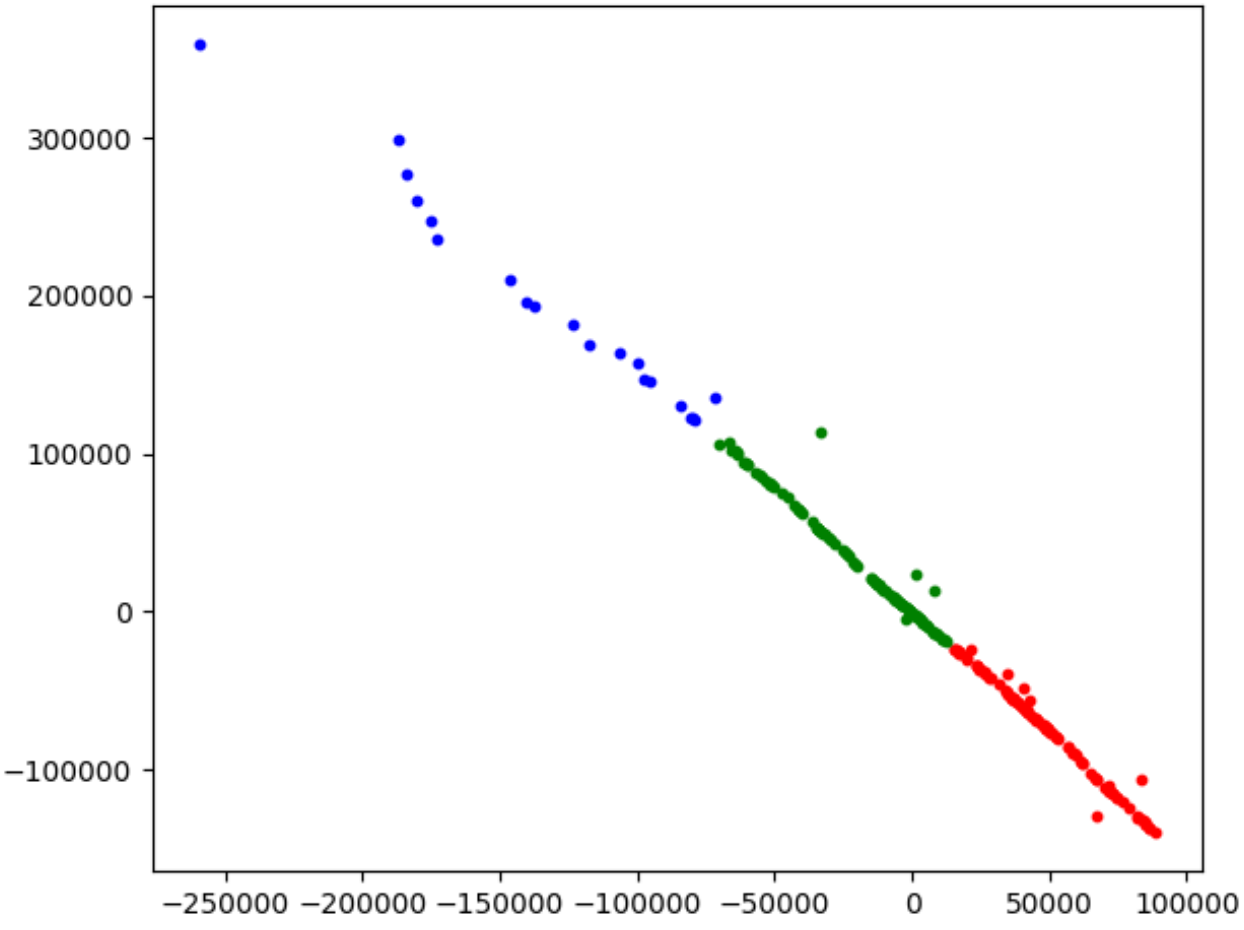
Далее, для определения числа кластеров строим график «Каменистая осыпь» («Локоть»).



По данному рисунку видна точка K = 3, в которой происходит резкое изменение графика (так называемый «Локоть»). Данная точка обозначает, количество кластеров, на которое имеет смысл делить имеющийся набор данных, так как при K < 3 данные в рамках одного кластера будут иметь большие различия, а при K > 3 данные в рамках двух разных кластеров будут иметь небольшие различия.

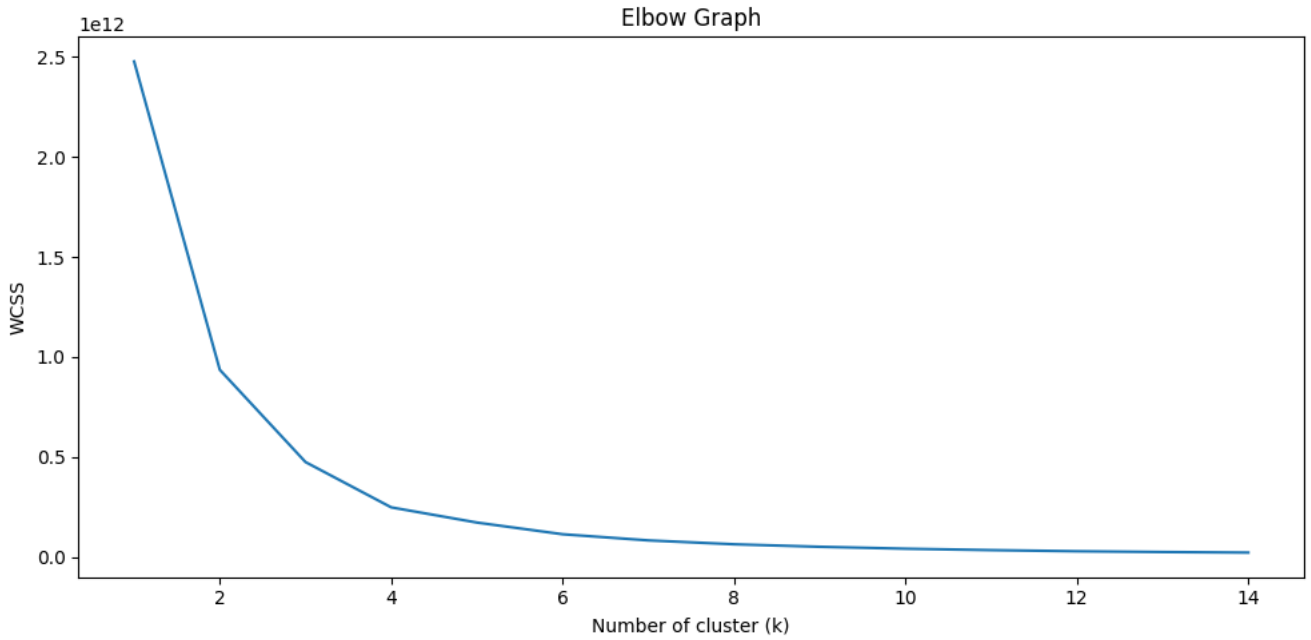
2) Метод K средних

Для построения графика метода K средних используются библиотеки sklearn.cluster.Kmeans (для получения результата метода) и sklearn.manifold.MDS (аналог метода cmdscale языка R). При помощи ресурсов, указанных выше, получаем график.



Разными цветами обозначены разные кластеры.

Количество кластеров равное 3 выбрано при помощи графика «Каменистая осыпь» («Локоть»).



Данный график немного отличается от графика, полученного при иерархической кластеризации, но в данном случае значение количества кластеров не различается.

Вывод

В итоге видно, что кластеры, полученные в результате иерархической кластеризации и кластеризации методом K средних немного различаются по своему составу, так как используются различные методы кластеризации, но те люди, которые находятся в рамках одного кластера, считаются похожими по заработной плате (в рамках определенного значения), в то время как люди, принадлежащие разным кластерам, считаются более отдаленными друг от друга вплане заработка.