**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра ИС**

**Отчет**

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

**Тема: «Кластеризации K-means»**

|  |  |
| --- | --- |
| Студентка гр. 2373 | Маркова З.А. |
| Преподаватель | Татчина Я.А. |

Санкт-Петербург 2024

**Цель работы**: познакомиться с методом кластеризации K-mean с помощью пакета sklearn.

1. **Краткое описание набора данных**

Для данной работы я решила поменять предыдущий датасет на датасет «Заработная плата по должности и стране».

Атрибуты датасета:

* Age – возраст. Тип данных числовой.
* Gender – пол. Тип данных стрококовый
* Education Level – уровень образования. (0 - High School, 1 - Bachelor Degree, 2 - Master Degree, 3 – Phd)
* Job Title – вид профессии. Тип данных стрококовый
* Years of Experience – стаж работы. Тип данных числовой.
* Salary – зарплата. Тип данных числовой.
* Race – раса. Тип данных стрококовый
* Senior - руководящая должность. (0 – не занимал, 1 – занимал).

Мною был добавлен новый атрибут Age of Employment – возраст трудоустройства. Age of Employment=Age - Years of Experience.

**2. Выбросы, пропуски, дубли**

В данном датасете было достаточно много дублей. Которые были убраны с помощью df.drop\_duplicates().

Пропусков в датасете не оказалось.

При поиске выбросов с помощью межквартильного диапазона, были найдены выбросы у атрибутов Age, Years of Experience, Age of Employment.

Однако, если мы посмотрим на гистограммы данных атрибутов, можно увидеть, что данные, выходящие за «норму» межквартильного диапазона, могут быть в реальной жизни, поэтому их нельзя считать ошибочными и можно оставить.

**3. Графики зависимости.**

Мною были построены 5 графиков зависимостей, также для каждого графика были выделены группы.

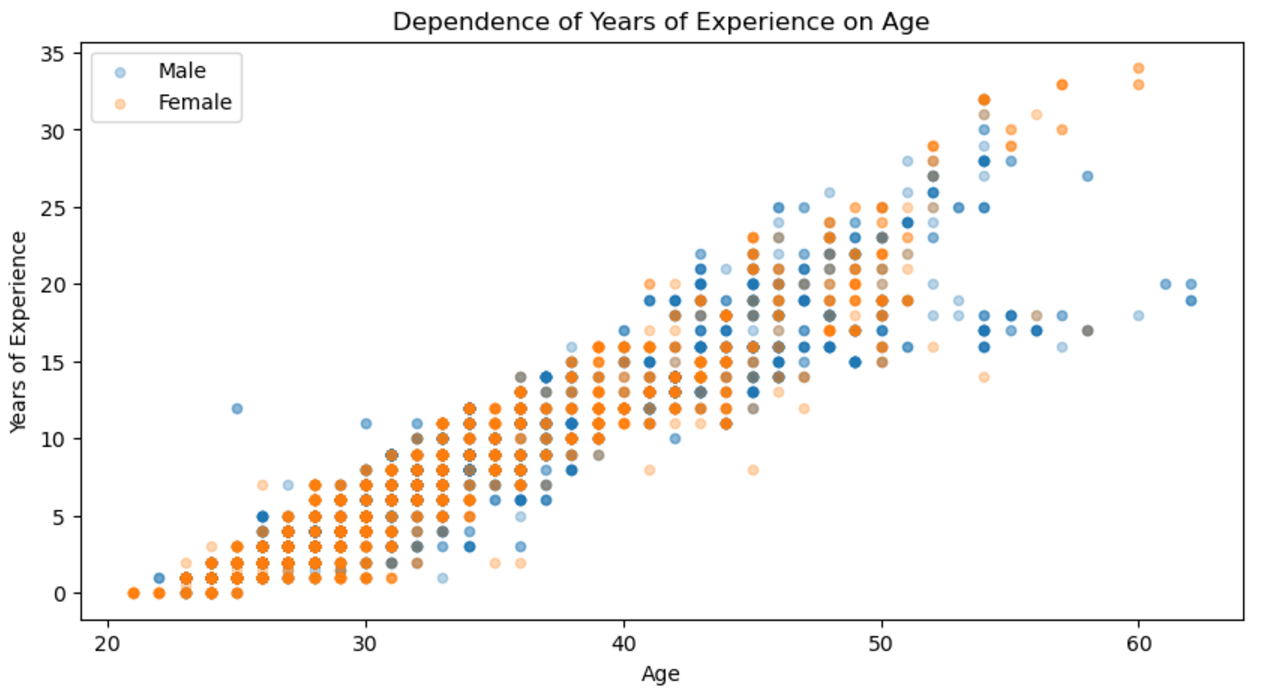
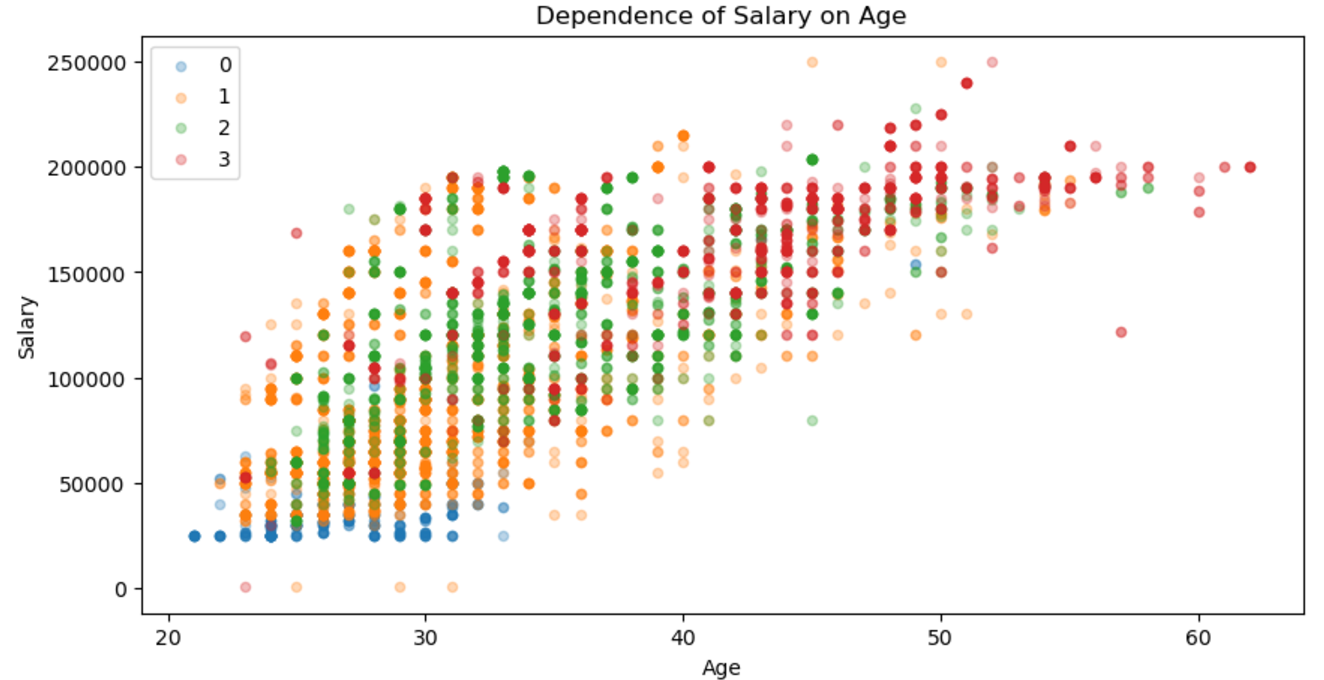
Рис.1 График зависимости Years of Experience от Age с отображением групп по полу

Рис.2 График зависимости Salary от Age с отображением групп по уровню образования

Рис.3 График зависимости Salary от Years of Experience с отображением групп по руководящей должности

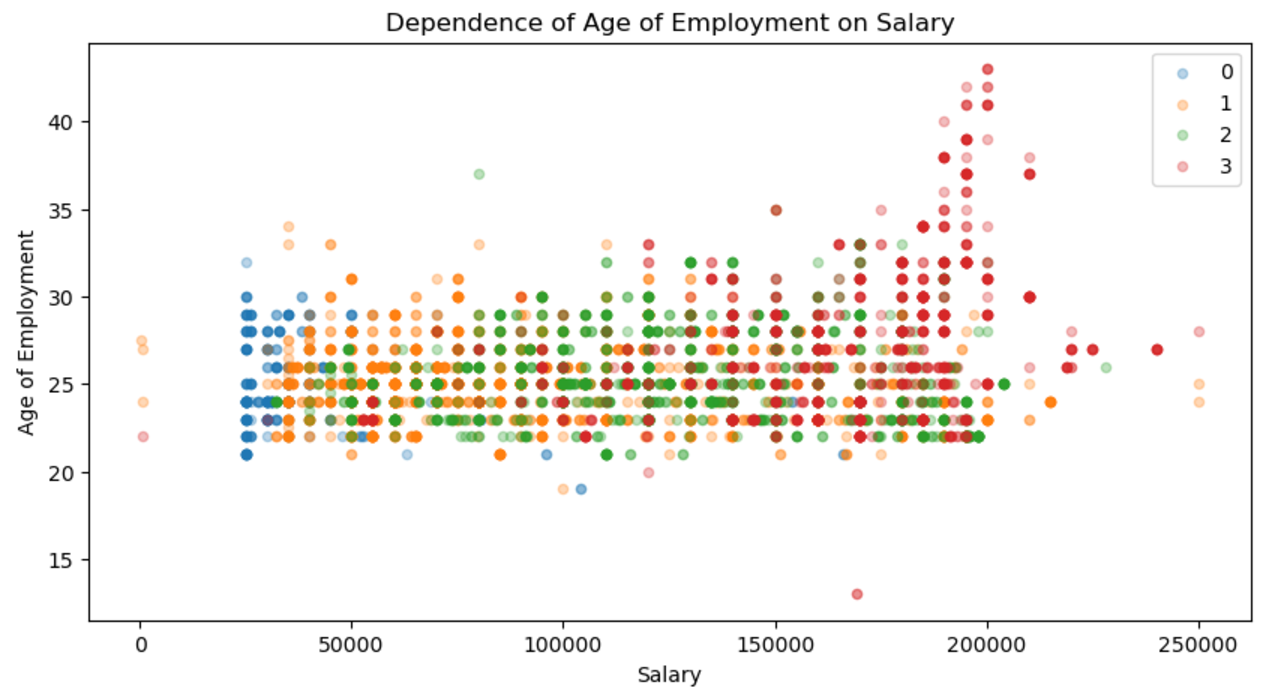
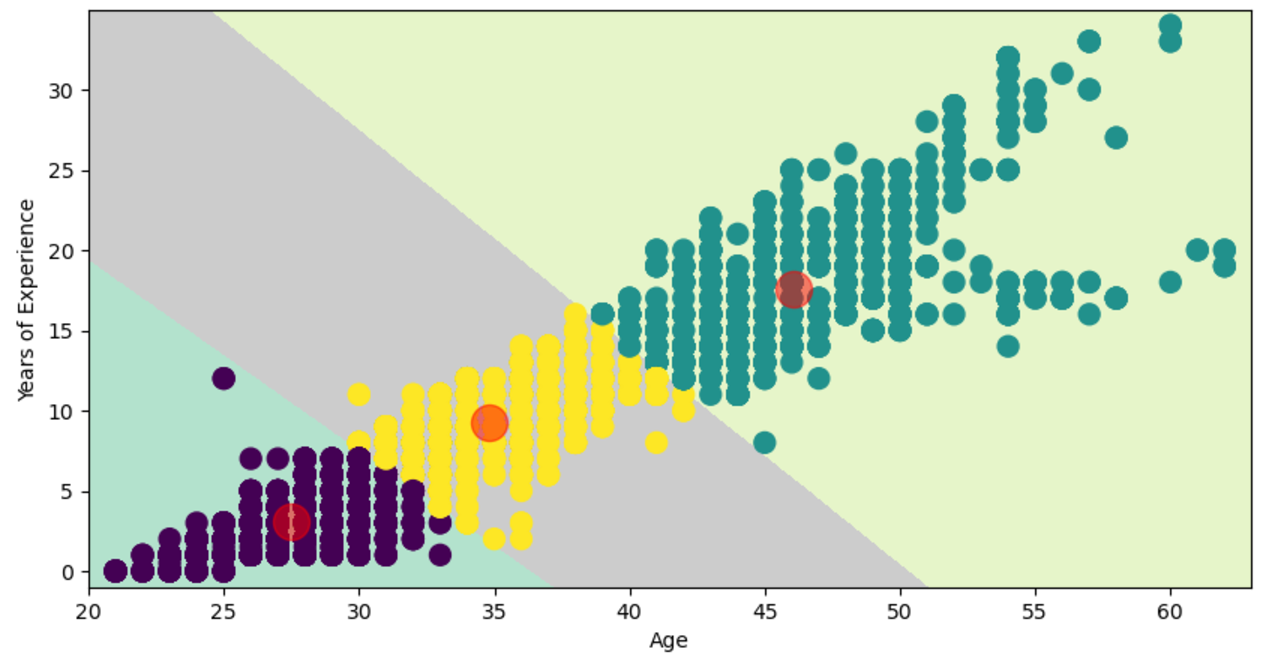
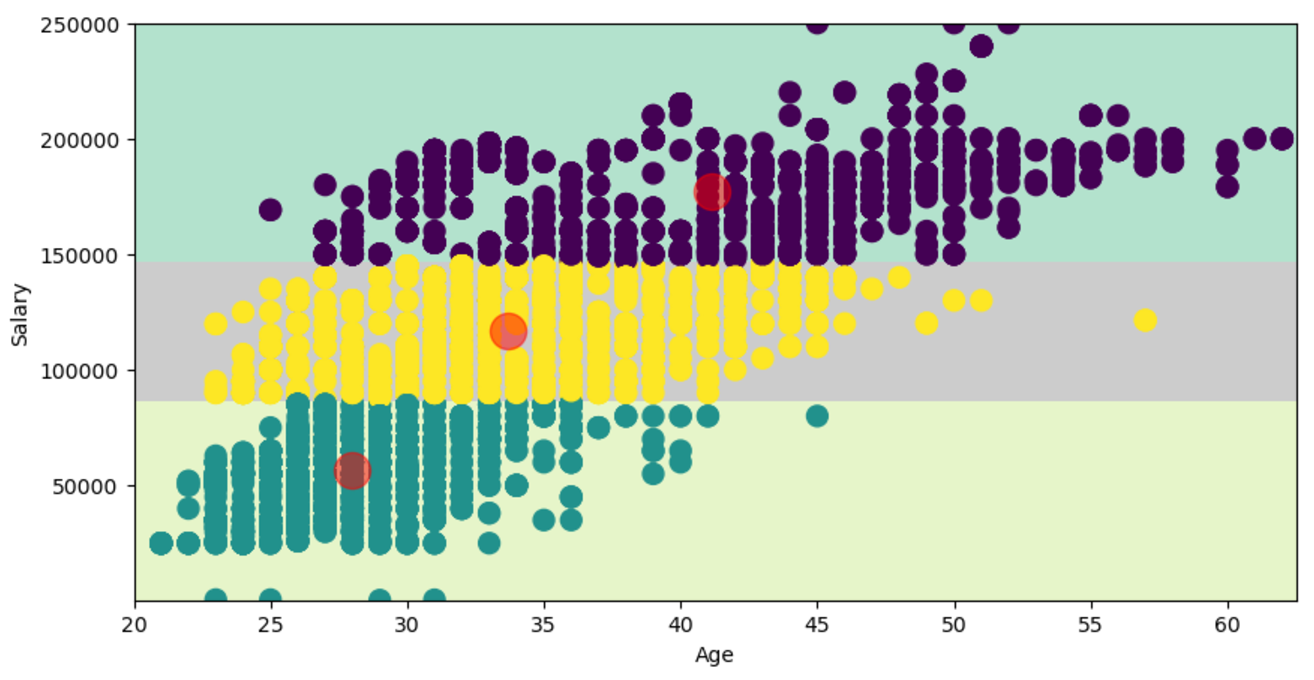
Рис.4 График зависимости Age of Employment от Salary с отображением групп по уровню образования

Рис.5 График зависимости Age of Employment от Age с отображением групп по полу

**4. Метод KMeans**

Сначала с помощью «метода локтя» нами было найдено оптимально количество кластеров для каждого графика. В результате было получено, что для каждого графика оптимальное количество кластеров – 3.

Метод KMeans разбивает множество элементов векторного пространства на заранее известное число кластеров k. 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.

Рис.6 Кластеризация для графика Years of Experience от Age

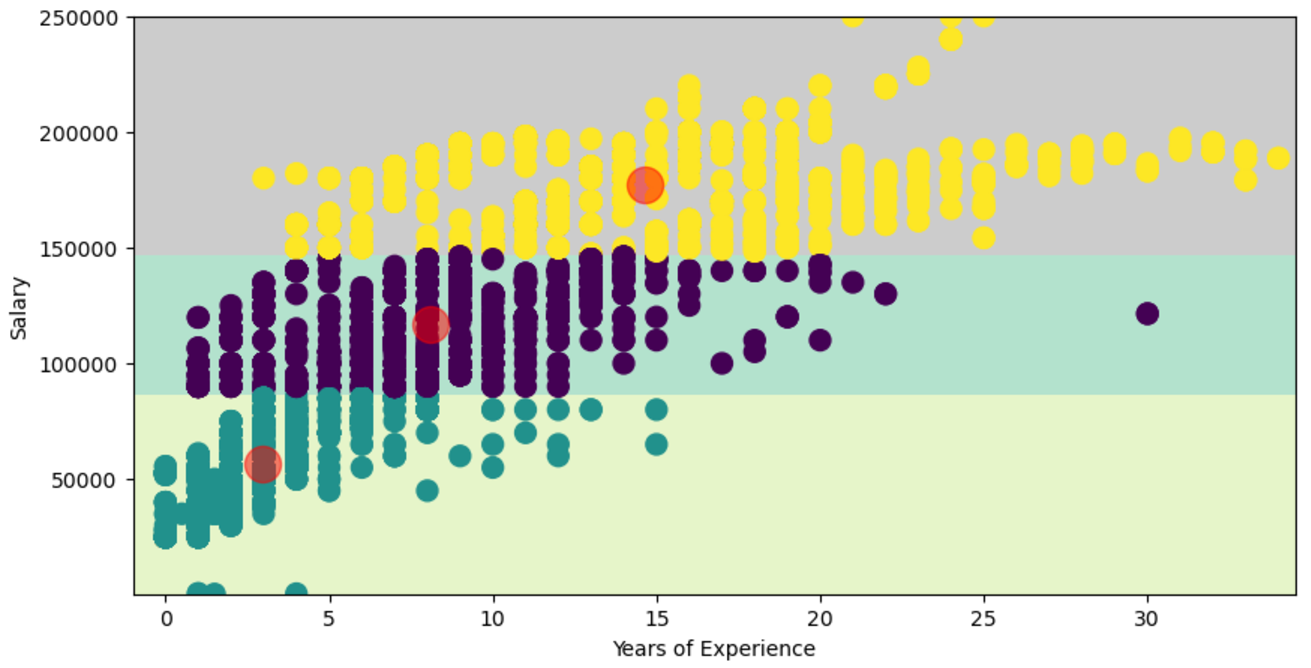
Рис.7 Кластеризация для графика Salary от Age

Рис.8 Кластеризация для графика Salary от Years of Experience

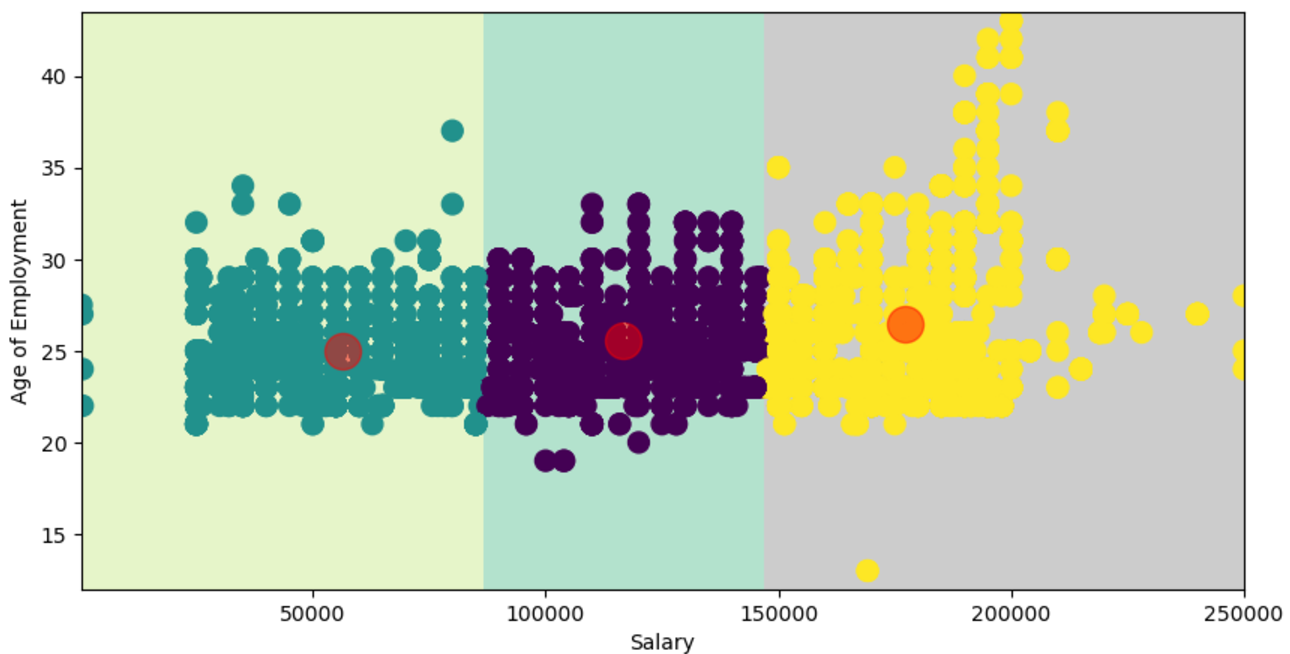
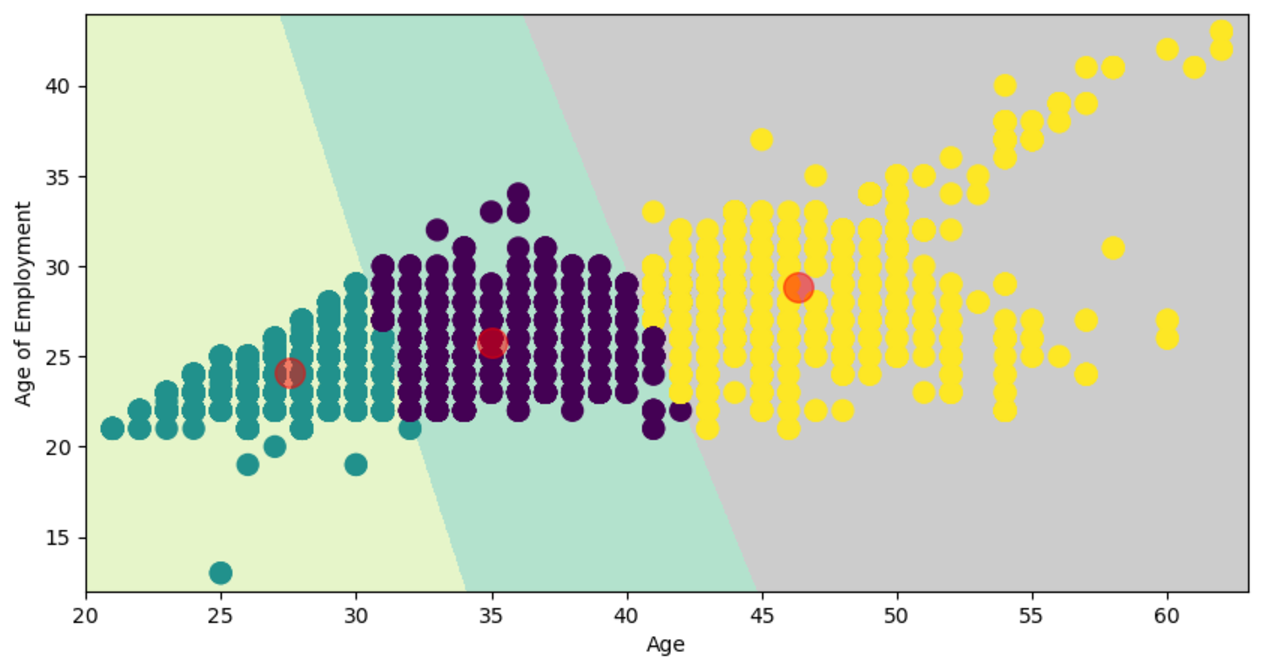
Рис.9 Кластеризация для графика Age of Employment от Salary

Рис.9 Кластеризация для графика Age of Employment от Age

**Вывод**

В данной лабораторной работе мы познакомились с методом кластеризации Kmeanы с помощью пакета sklearn.

Был обработан датасет «Заработная плата по должности и стране». Из него были убраны дубли и найдены выбросы. Также был добавлен новый атрибут Age of Employment.

Для графиков зависимостей были выделены группы, из которых видно, что в данном датасете:

* Взяты данные для женщин более молодого возраста, в отличие от мужчин
* Зарплата зависит от возраста и уровня образования
* Люди со средним стажем и со средней зарплатой чаще занимали руководящую должность
* Люди с разным уровнем образования, устраиваются на работу в любом возрасте
* Женщины устраивались на работу в более молодом возрасте

Для каждого из данных графиков было найдено оптимальные количество кластеров. Во всех случаях получилось 3. Были построены графики кластеризации с помощью метода KMeans.