МИНОБРНАУКИ РОССИИ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ

ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ «ЛЭТИ»

ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)

Кафедра вычислительной техники

ОТЧЕТ

по лабораторной работе №3

по дисциплине «Машинное обучение»

Тема: «Кластеризация»

Студентка гр.0305 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Петракова М. А.

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Гатауллин Р. И.

Санкт – Петербург

2023

**Задание**

1. Обучение моделей и подбор параметров (где применимо):

a.     метод К-средних

b.     DBSCAN

c.     Иерархическая кластеризация

2. Оценка моделей

а. Экспертная оценка

b. Сравнение разбиения на классы с помощью кластеризации с реальными.

c. Визуализация предсказанных значений

**Теоретическая часть**

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Один из наиболее популярных методов кластеризации — это метод K-средних (K-means). Основная идея метода — итеративное повторение двух шагов:

1. распределение объектов выборки по кластерам;
2. пересчёт центров кластеров.

В начале работы алгоритма выбираются K случайных центров в пространстве признаков. Каждый объект выборки относят к тому кластеру, к центру которого объект оказался ближе. Далее центры кластеров пересчитывают как среднее арифметическое векторов признаков всех вошедших в этот кластер объектов (то есть центр масс кластера). Как только мы обновили центры кластеров, объекты заново перераспределяются по ним, а затем можно снова уточнить положение центров. Процесс продолжается до тех пор, пока центры кластеров не перестанут меняться.

Алгоритм DBSCAN (Density-based spatial clustering of applications with noise) развивает идею кластеризации с помощью выделения связных компонент.

Все объекты выборки делятся на три типа: внутренние / основные точки (core points), граничные (border points) и шумовые точки (noise points). К основным относятся точки, в окрестности которых больше объектов выборки. К граничным — точки, в окрестности которых есть основные, но общее количество точек в окрестности меньше. Шумовыми называют точки, в окрестности которых нет основных точек и в целом содержится менее

объектов выборки.

Алгоритм кластеризации выглядит следующим образом:

1. Шумовые точки убираются из рассмотрения и не приписываются ни к какому кластеру.
2. Основные точки, у которых есть общая окрестность, соединяются ребром.
3. В полученном графе выделяются компоненты связности.
4. Каждая граничная точка относится к тому кластеру, в который попала ближайшая к ней основная точка.

Другой классический метод кластеризации — это иерархическая агломеративная кластеризация. Агломеративные алгоритмы начинают с небольших кластеров (обычно с кластеров, состоящих из одного объекта) и постепенно объединяют их в кластеры побольше.

Сам алгоритм:

1. Создаём столько кластеров, сколько у нас объектов в выборке, каждый объект — в своём отдельном кластере.
2. Повторяем итеративно слияние двух ближайших кластеров, пока не выполнится критерий останова.

**Выполнение работы**

Предварительно была произведена фильтрация данных: в наборе имеются только ненулевые данные, оставлены столбцы: 

Кластеризация проводится по первым 5 признакам. Они нормированы и приведены к числовому типу.

Набор данных, на котором проводится кластеризация приведен на рис.1.

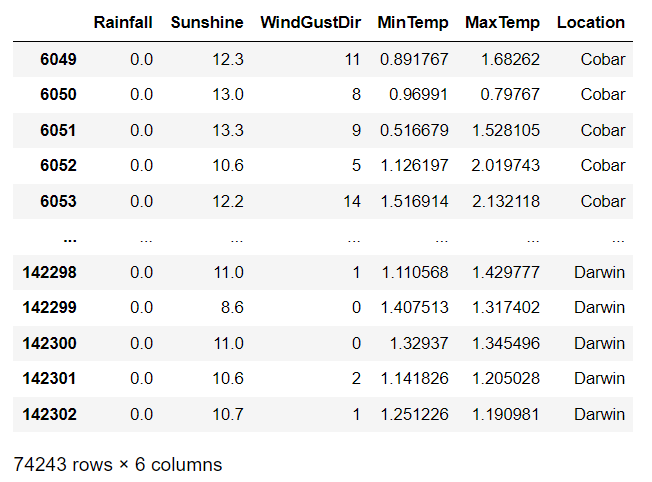
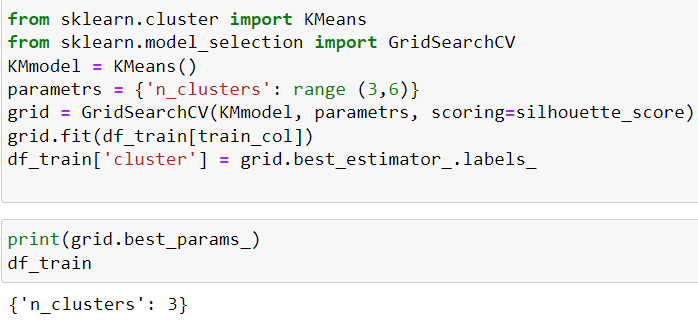


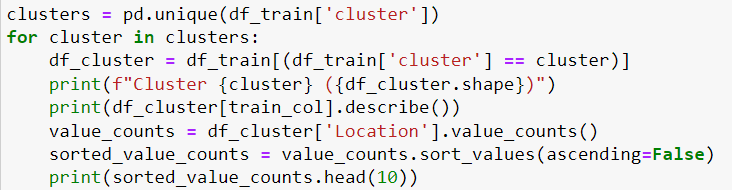
Рисунок 1

1. **Kmeans**

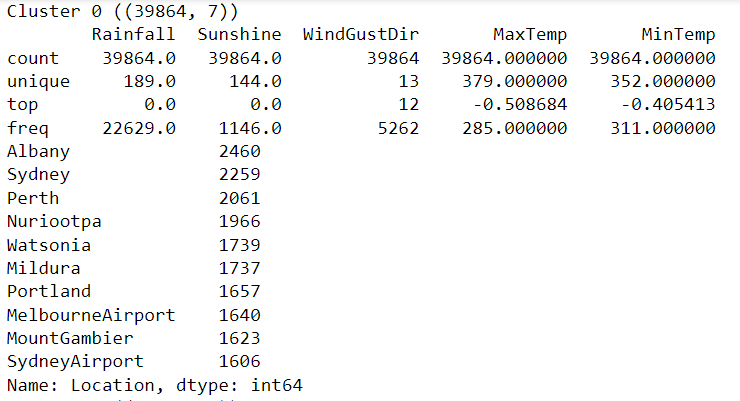
Выбор наилучшего числа кластеров посредством GridSearch: 3 кластера. Обучение модели.

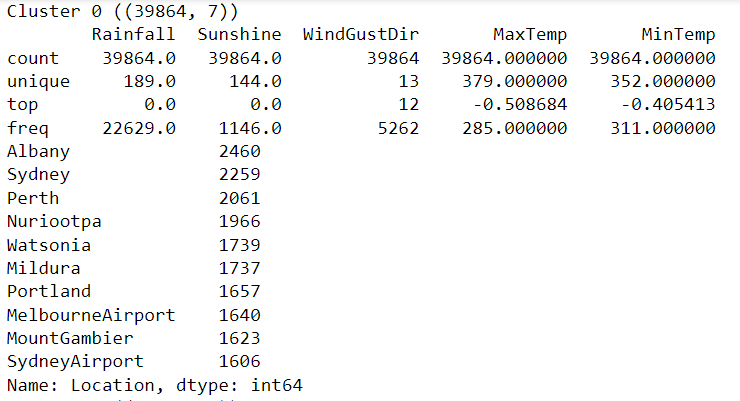


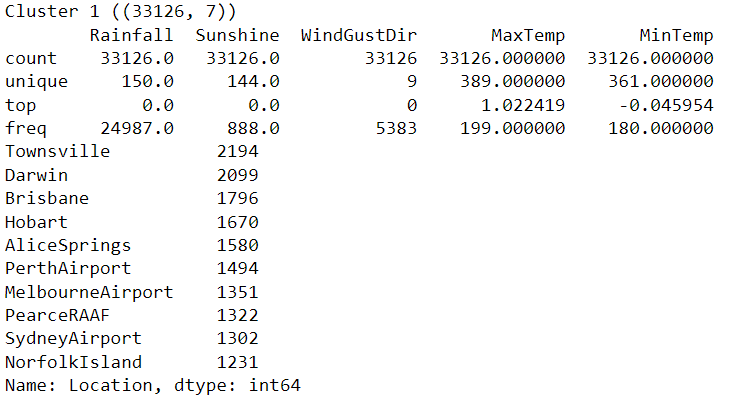
Вывод информации о кластерах.

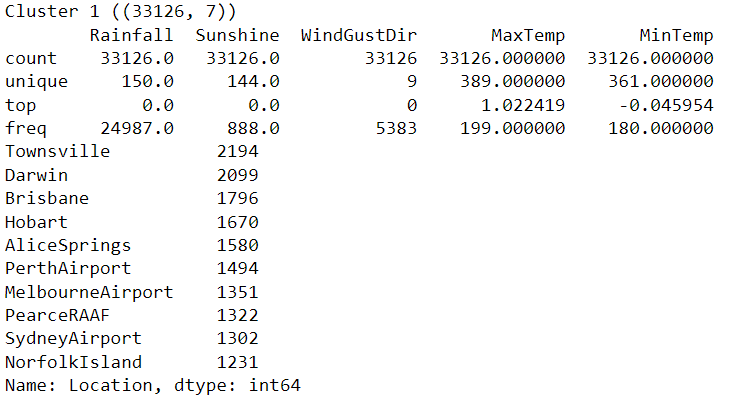


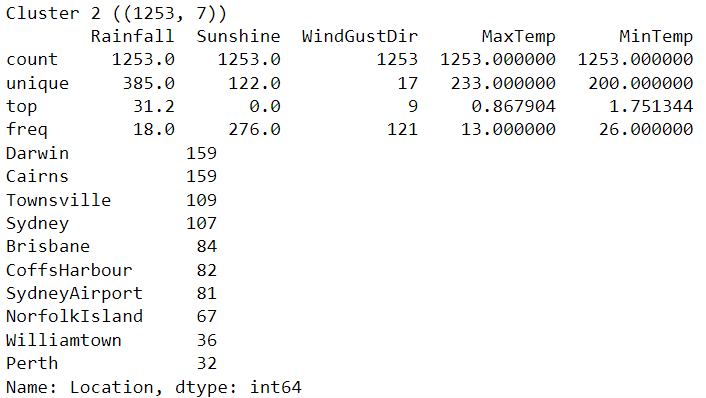
Кластер 0

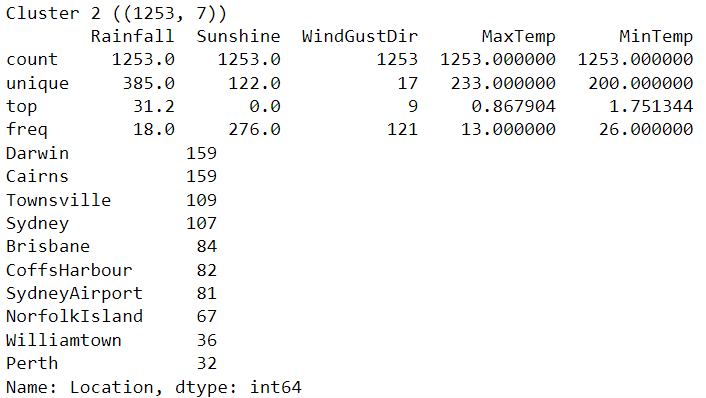




Кластер 1

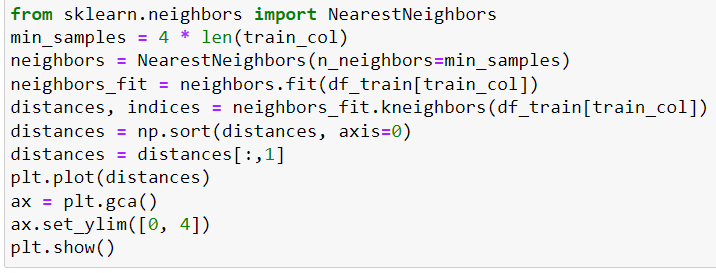


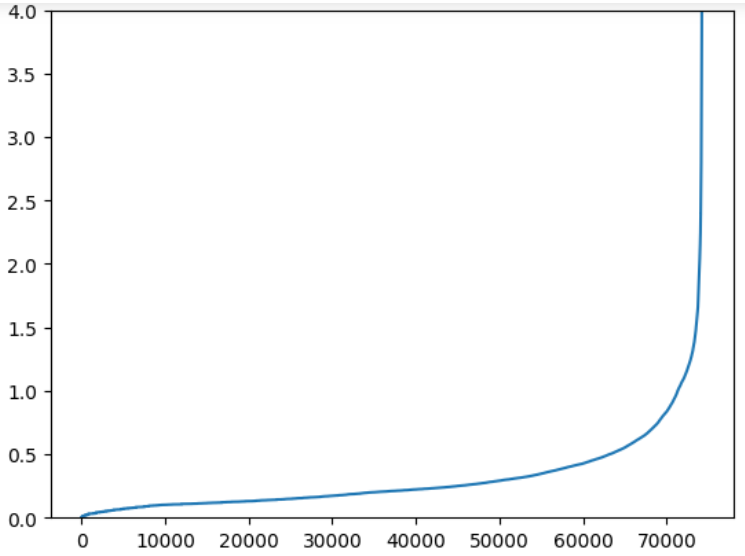
Кластер 2



1. **DBSCAN**

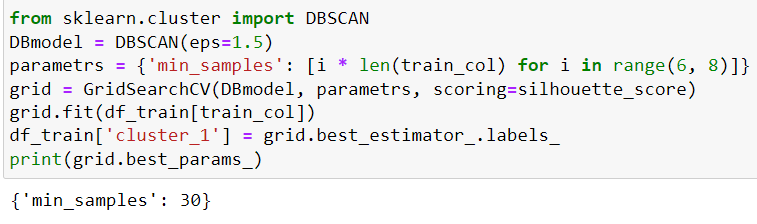
Выбор расстояния Эпсилон.



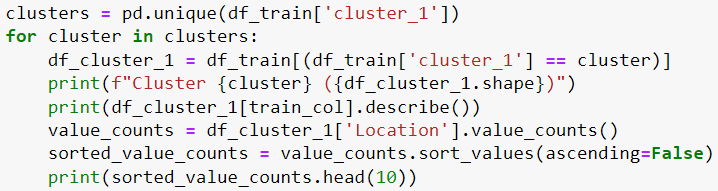


Эпсилон возьмем равной 1.25.

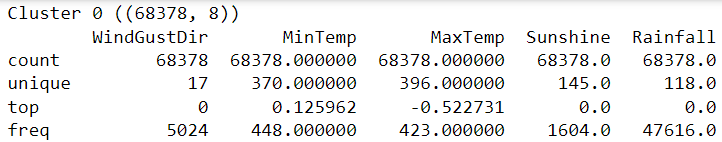
Выбор наилучшего числа экземпляров в ядре посредством GridSearch: 30. Обучение модели.

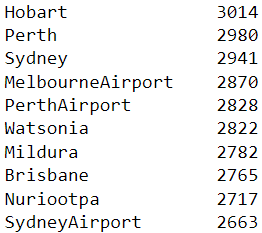


Вывод информации о кластерах.

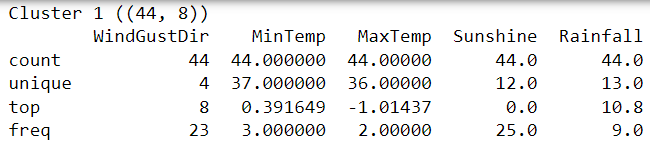


Кластер 0



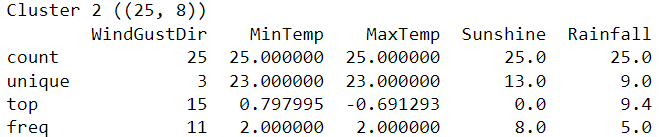
 

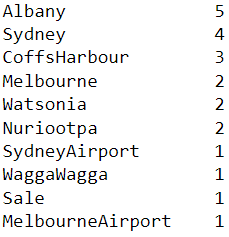
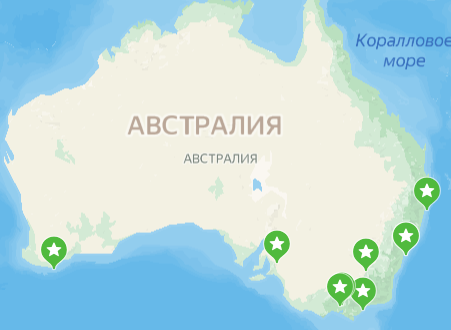
Кластер 1



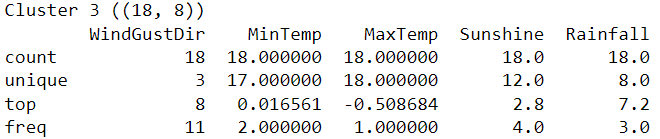
 

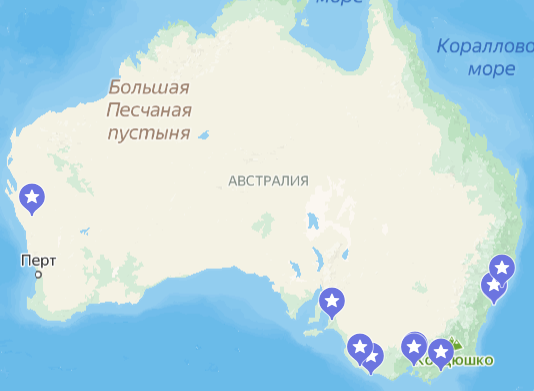
Кластер 2



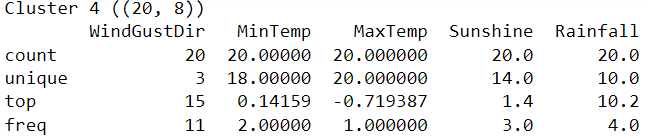
 

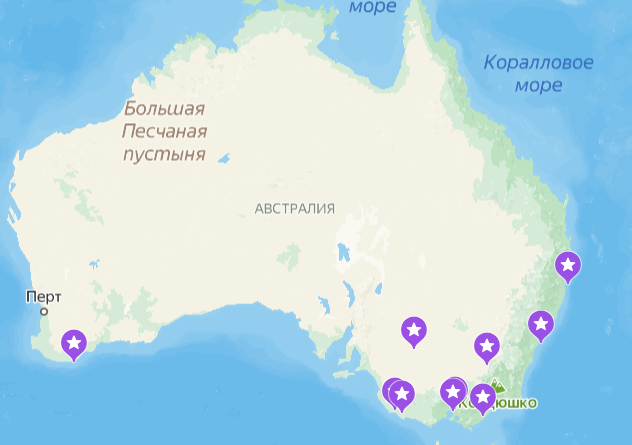
Кластер 3



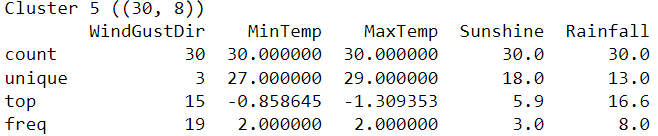
 

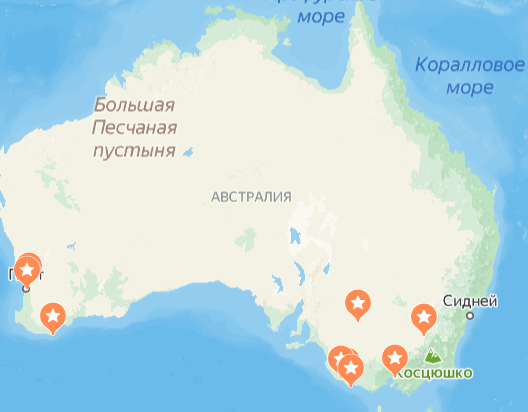
Кластер 4



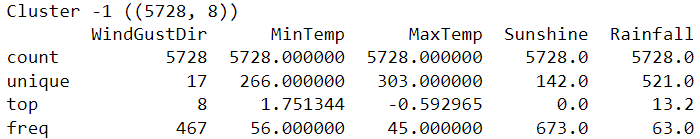
 

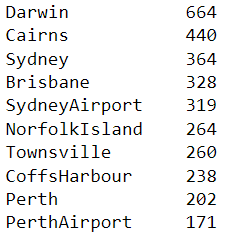
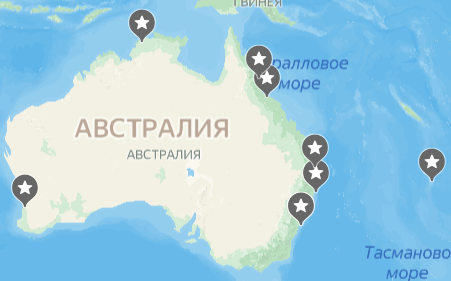
Кластер 5



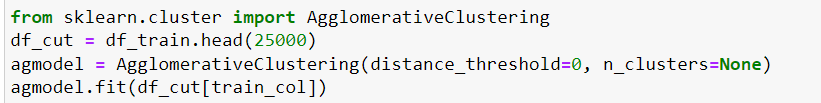
Кластер -1 (выбросы)



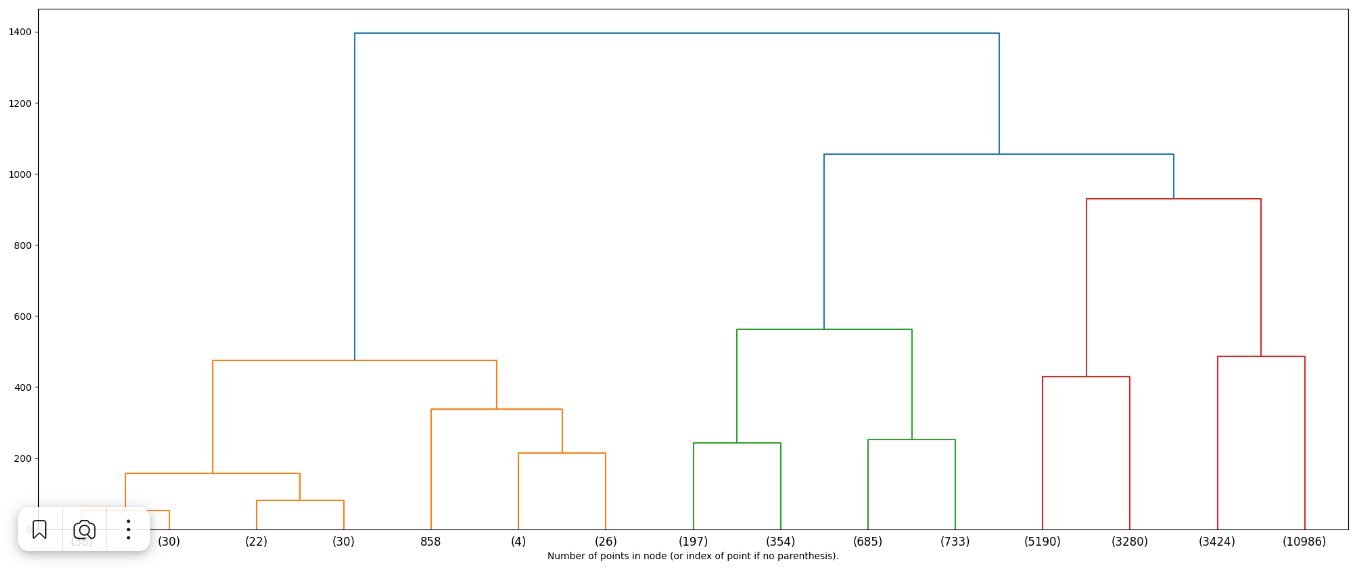
 

1. Иерархическая кластеризация

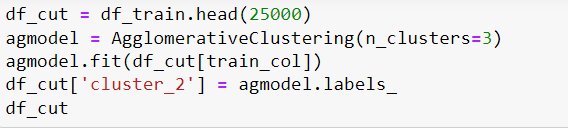
Поиск оптимального числа кластеров.



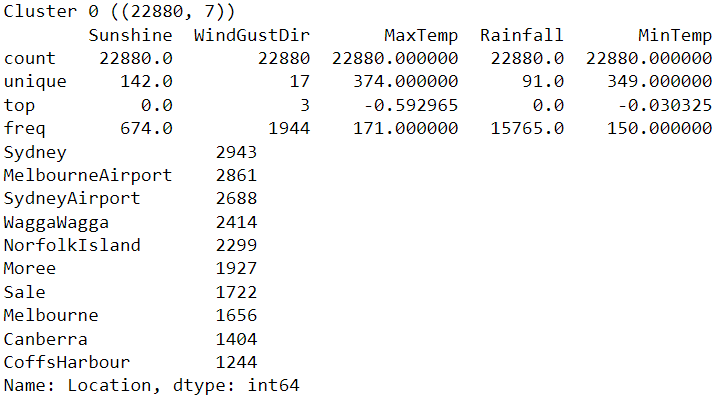
Дендрограмма.



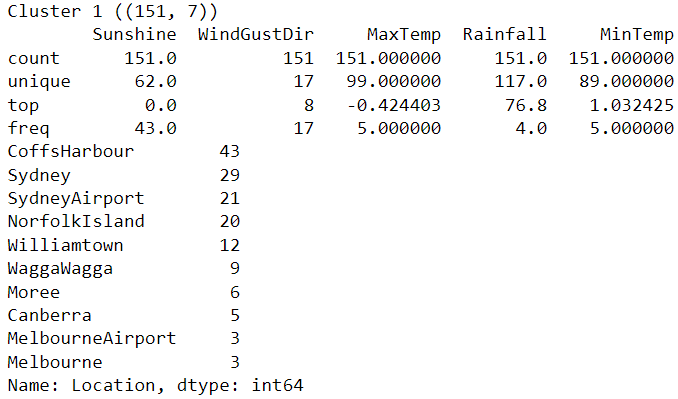
Обучение модели (3 кластера).

  
Результат.

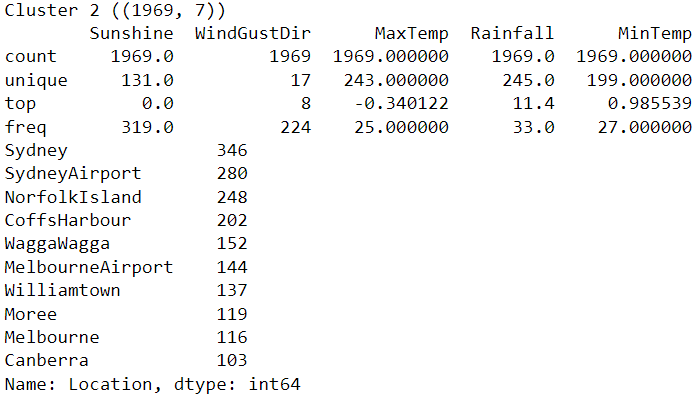
Кластер 0



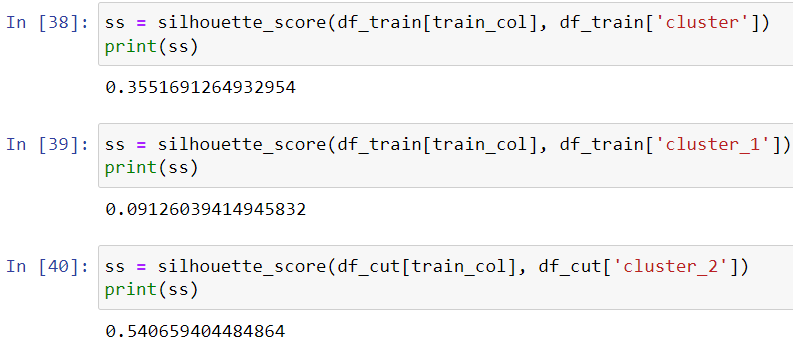
Кластер 1



Кластер 2



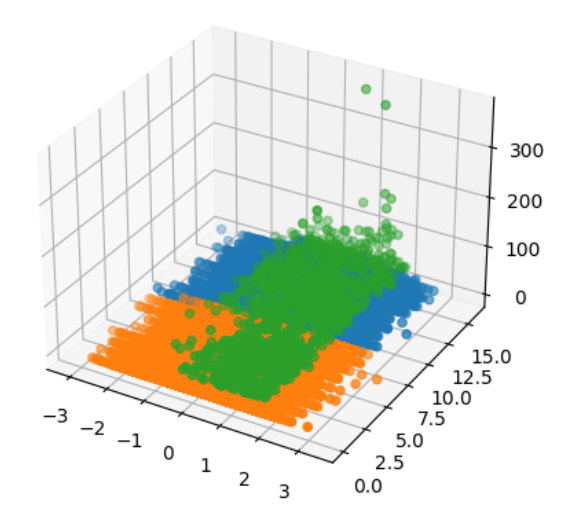
Оценка моделей.

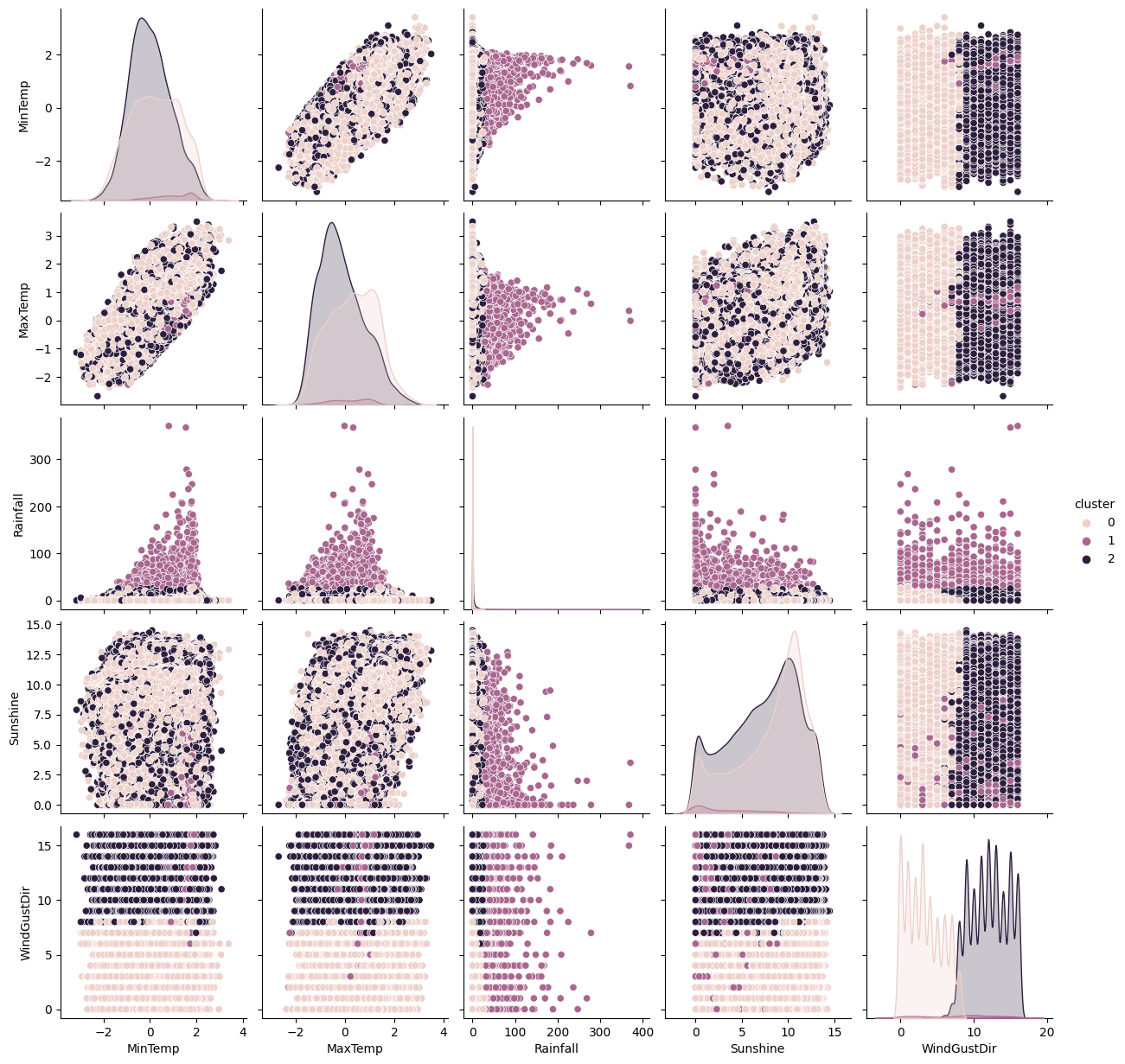


Наилучшие результаты – у иерархической.

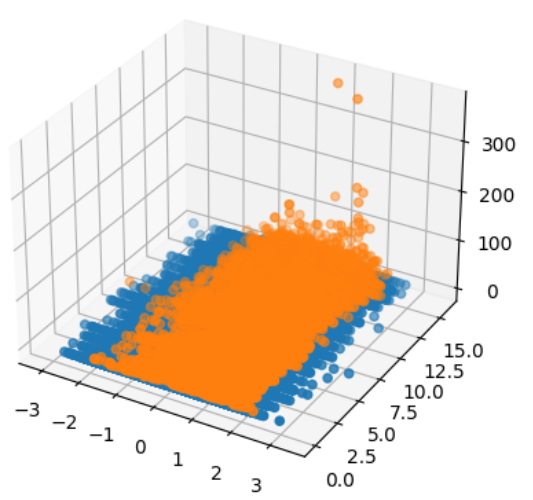
Визуализация кластеров.

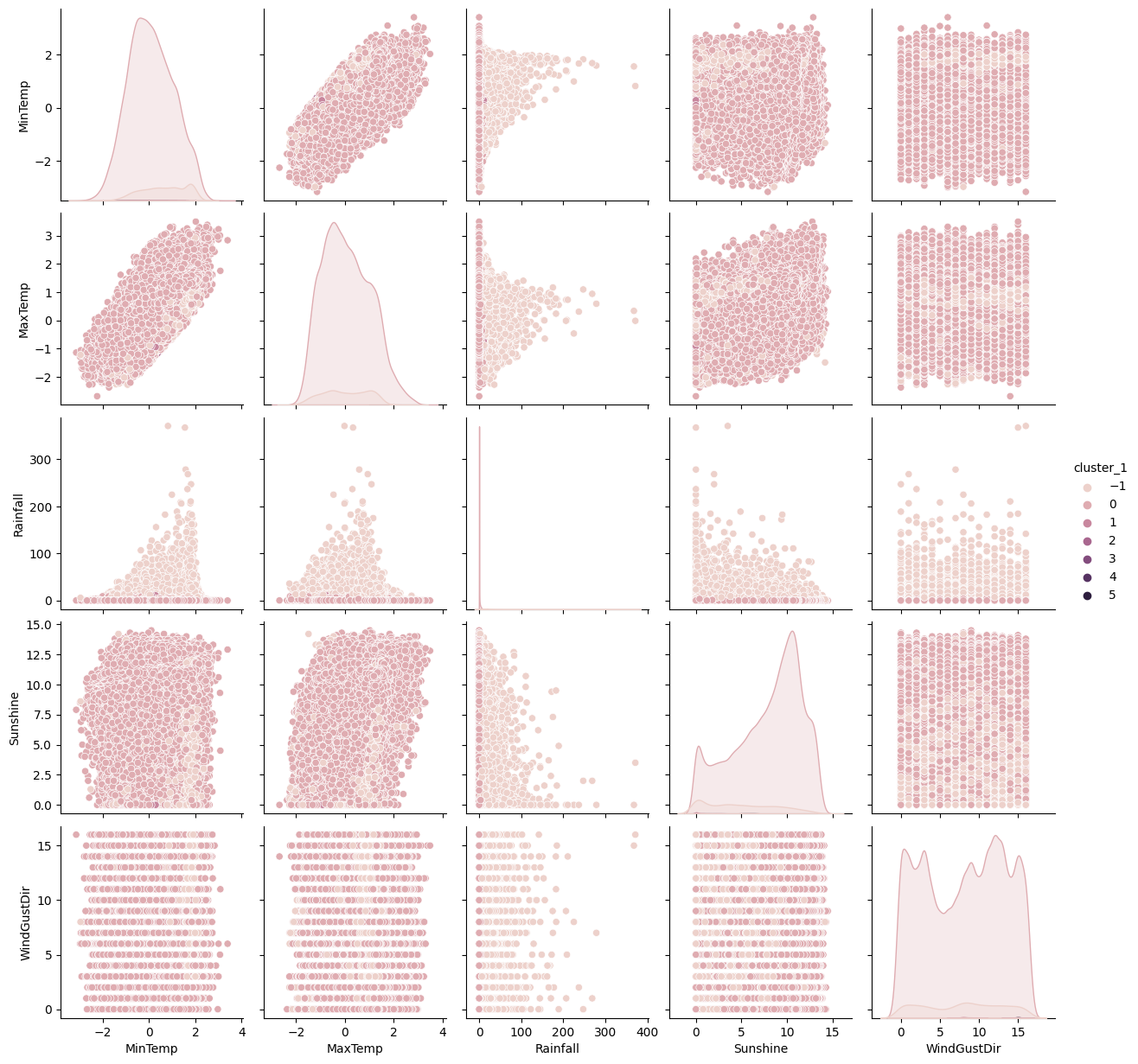
1. метод К-средних



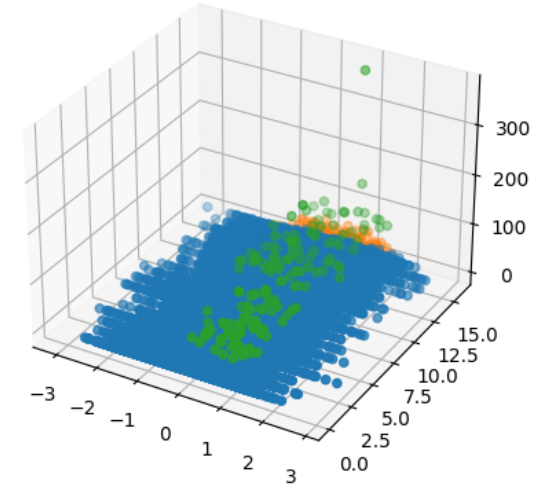


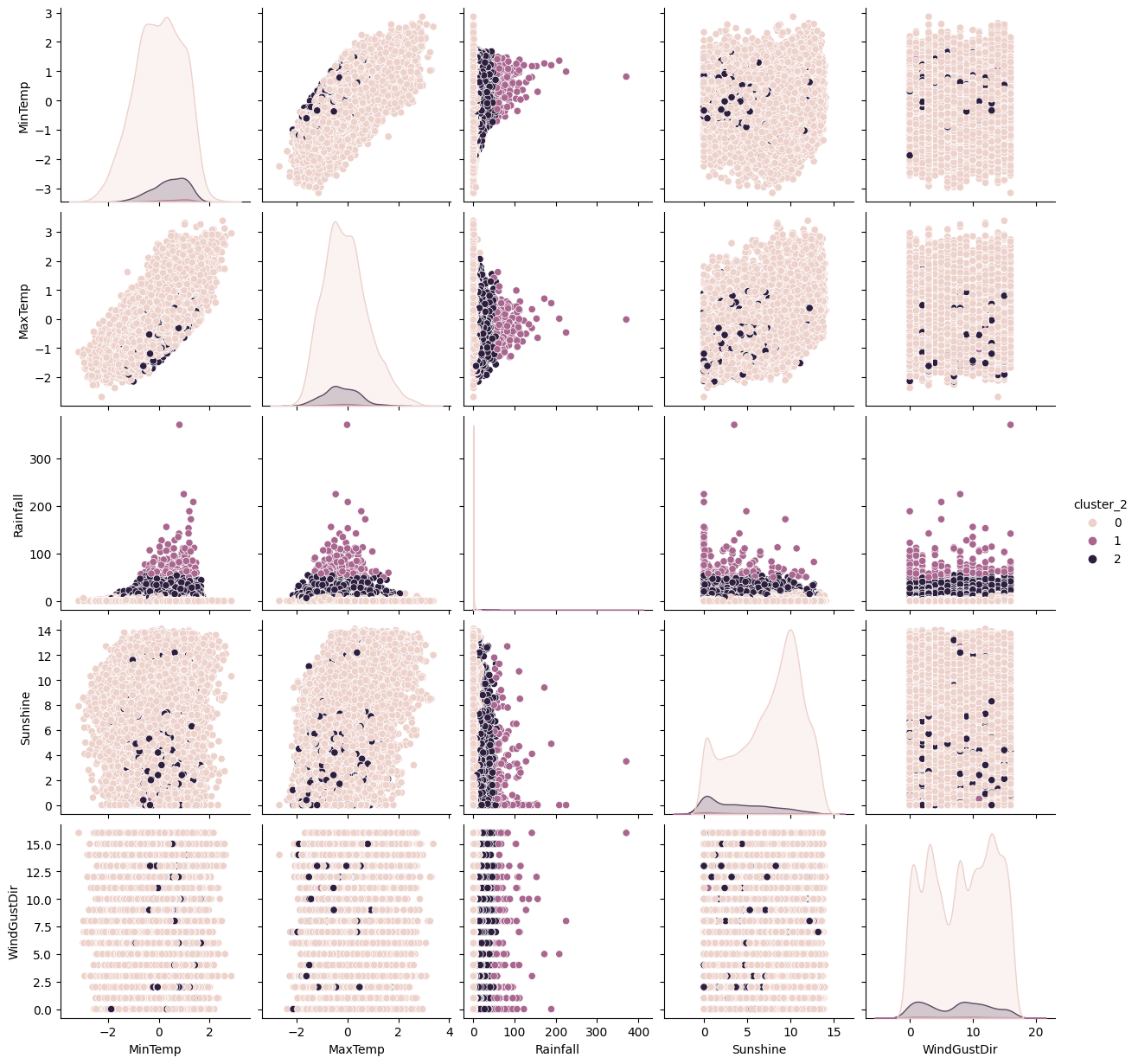
1. DBSCAN





1. Иерархическая кластеризация





**Выводы.**

Данные были кластеризованы тремя методами: К-средних, DBSCAN и иерархическим методом. Кластеры отличаются по размеру. Кластеры, полученные разными моделями, отличаются. Разбиение данных производилось на основе 5 параметров: по результатам можно увидеть наиболее частое значение каждого из параметров в кластере и его частоту. По этому видно, какой параметр был решающим при определении кластера для объекта.

Кластеры объединяют объекты (строки в DataFrame), хранящие значения о погодных условиях и местоположении. По визуализации кластеров видно, что места были объединены по разным параметрам: наиболее дождливые регионы, наиболее солнечные и т.д.

В силу большого объема данных запустить обучение модели иерархической кластеризации не удалось, она обучается на 30% данных, поэтому полное сравнение моделей не будет точным.

Видно, что наибольшее влияние на принадлежность кластеру модель К-средних посчитала направление ветра. Иерархическая модель – количество осадков.