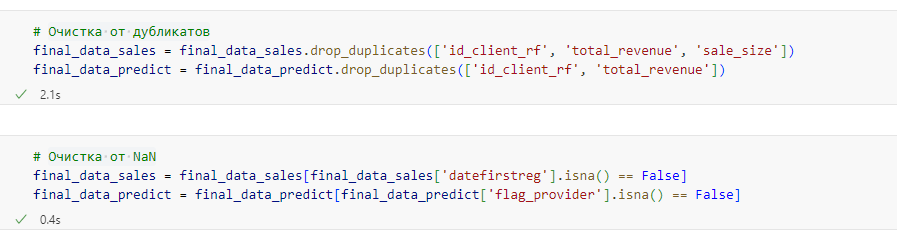
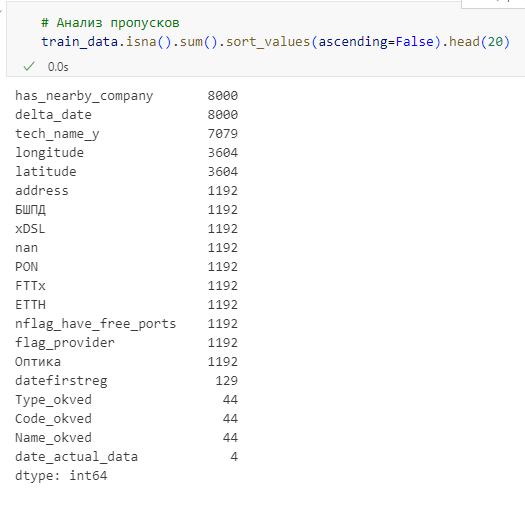
Вопросы к скрипту:

1. Зачем удалять дубли. Решили вроде что филиалы по одному клиенту не будем удалять:

**Комментарий**: я тогда особо не вникала, мне нужно было модель попробовать запустить. Сейчас буду пробовать и так, и так.

1. В train\_data остались NaN. Насколько это критично?

**Комментарий**: К сожалению, критично. Алгоритмы ML не работают с NaN.  


1. Оставляем только первые 2 цифры Code\_okved – done

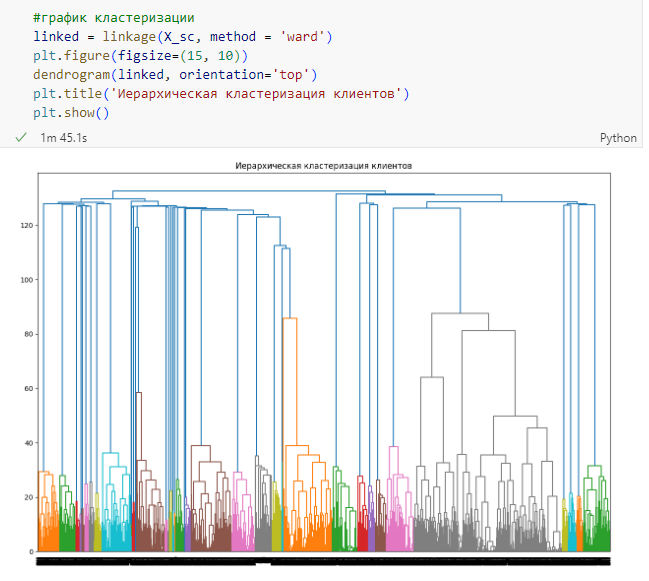
**Комментарий:** Отлично

1. У нас в данных по клиентам нет ИНН. Поэтому подтянуть внешние источники не понял, по какому ключу.

**Комментарий:** id – это и есть ИНН

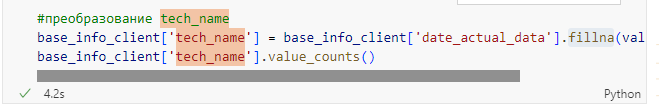
1. Кластеризация по дендраграммам по методу Варда вроде бы используется, когда небольшое кол-во кластеров:

**Комментарий:** это просто визуализация, можно и другое попробовать, лейблы записаны в другом месте.



1. Для чего в base\_info\_client создается поле 'tech\_name' записывается date\_actual\_data? Или это опечатка?

**Комментарий:** да, это опечатка



1. Касательно Силуэта не был импортирован метод:  
   from sklearn.metrics import silhouette\_score

Я вставил еще другие методы определения числа кластеров. Но я пока не силен в этом анализе, может тебе поможет.

**Комментарий:** там импортирован весь модуль метрик sklearn.metrics, поэтому и странно, почему он у меня не работал. Попробую отдельно испортировать.