Подготовка данных.

Отфильтрованы поля
l_col = ['first_decision_state', # одно значение
'ФИО клиента',
'id клиента',
'ИНН работодателя',
'Сделка дошла до Андерайтреа', # одно значение
'СФ андерайтера', # после цели
'Комментарии андерайтера', # после цели
'Коды отказа', # после цели
'Описание кодов отказа', # дублирует код отказа
'Коды системы', # после цели - заполняются Андерами.
'Описание кодов', # дублирует код системы
'СФ системы', # Системная проверка - это проверки по которым сделка ушла в отказ. Они могут установится как до решения так и после решения Андерайтера
'Системная проверка давшая СФ', # Системная проверка - это проверки по которым сделка ушла в отказ. Они могут установится как до решения так и после решения Андерайтера
'Статус заявки', # после цели
'Решение по заявке', # после цели
'ФИО АНД принявшего последнее решение',
'Дата создания заявки', # дублирует Дата и Время создания заявки
'Подтвержденная сумма кредита' # после цели
]

Созданы поля:
1) разница между Доход клиента и Подтвержденным доходом клиента
2) час создания заявки
3) месяц создания заявки
4) возраст клиента

Проведена вставка пропущенных значений для полей имеющих меньше 50% пропущенных значений, путем предсказания по другим полям. Кодирование категориальных данных с количеством категорий до 10 к "one hot" кодированию, более десяти к кодированию метками.

Для кластерного анализа: проведено шкалирование стандартизацией. Инженериг признаков (операция суммы и деления) показал себя плохо в корреляционном анализе, в кластерном дал слишком большую задержку и выбросы, и применился только для экспримента.

Количество объектов – 13118, рабочих признаков 23. После стандартного удаляения выбросов 12968

Корреляционный анализ.

Коэффициент линейной корреляции Пирсона

0.598
Доход клиента
Скоринговый балл ОКБ, основной скоринг бюро

0.361061
Анкетный скоринг
Возраст клиента

0.352448
Эквифакс 4Score
Возраст клиента

0.306542
Запрошенная сумма кредита
Подтвержденный доход клиента

0.285754
Эквифакс 4Score
Анкетный скоринг

-0.273370
Мегафон
Месяц создания заявки

0.223299
Оценка кредитной истории Эквифакс_КИ отсутствует
Мегафон

Коэффициент нелинейной корреляции Спирмена показывает то же самое.

Итог: Новых полезных зависемойсте найдено не было.

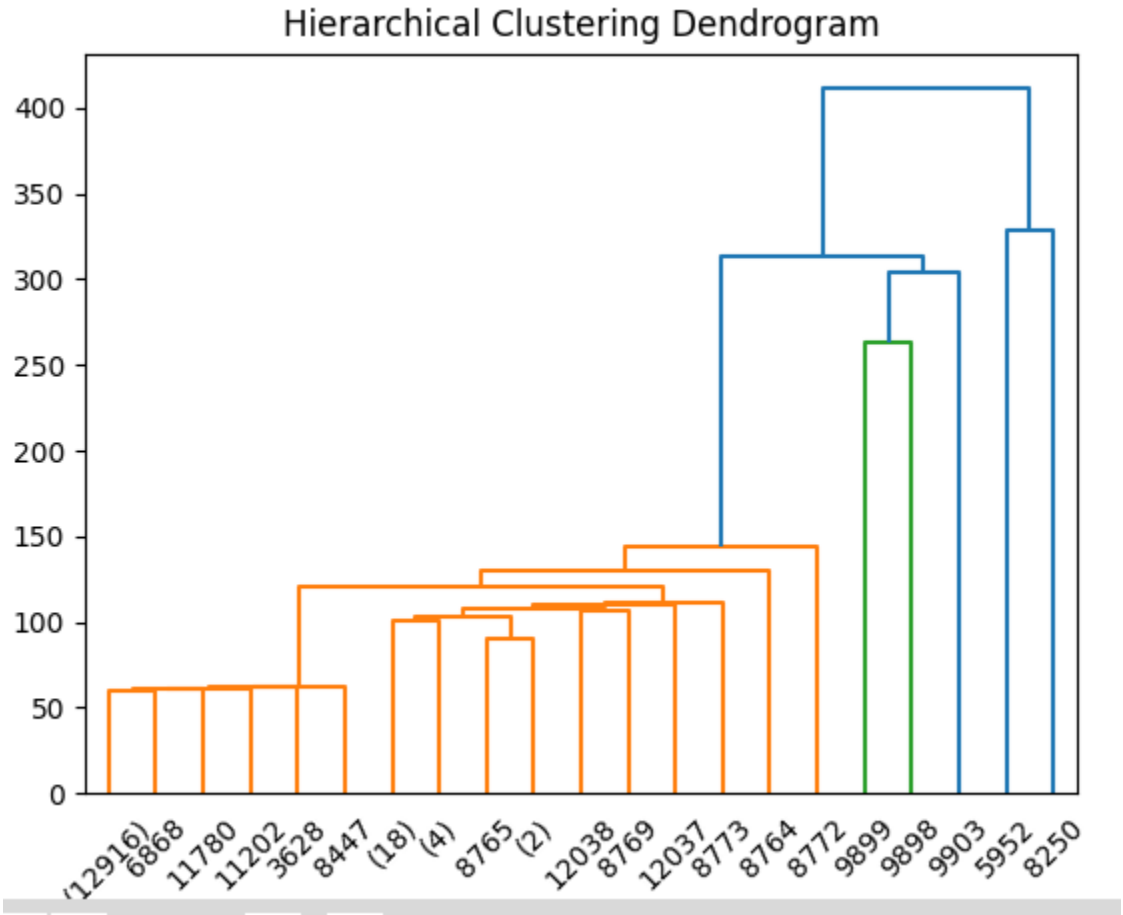
Иерархический кластерный анализ

Проведем иерархический кластерный анализ петум итераций от этапа когда каждая запись это класер к завершению когда все кластеры сливаются в один. (Агломеративный метод)

Поиск плотных шаровых скоплений

affinity='euclidean'
linkage='ward'
Метод Варда находит плотные шаровые скопления, в качестве метрики используется евклидово расстояние.

С инженерингом признаков.

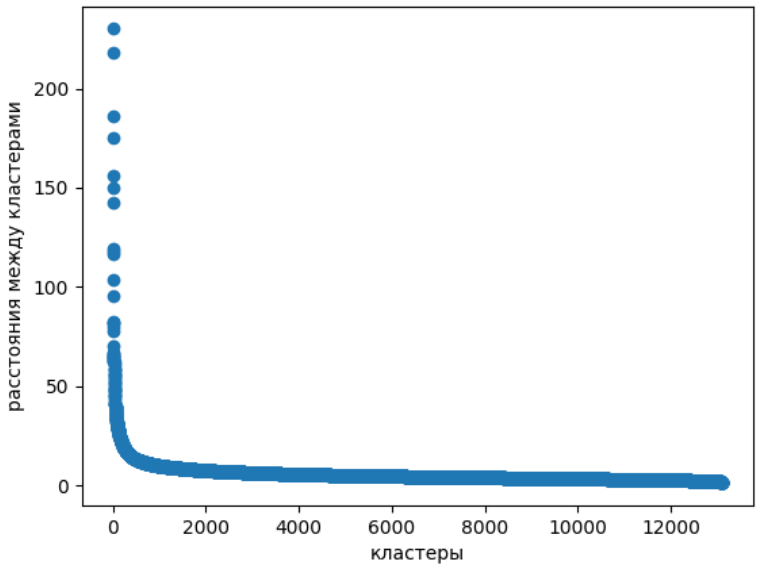


Дендрограммма показыват нам один растущий кластер с большим количеством выбросов.

Без инженеринга признаков

Евклидова метрика, ward методом

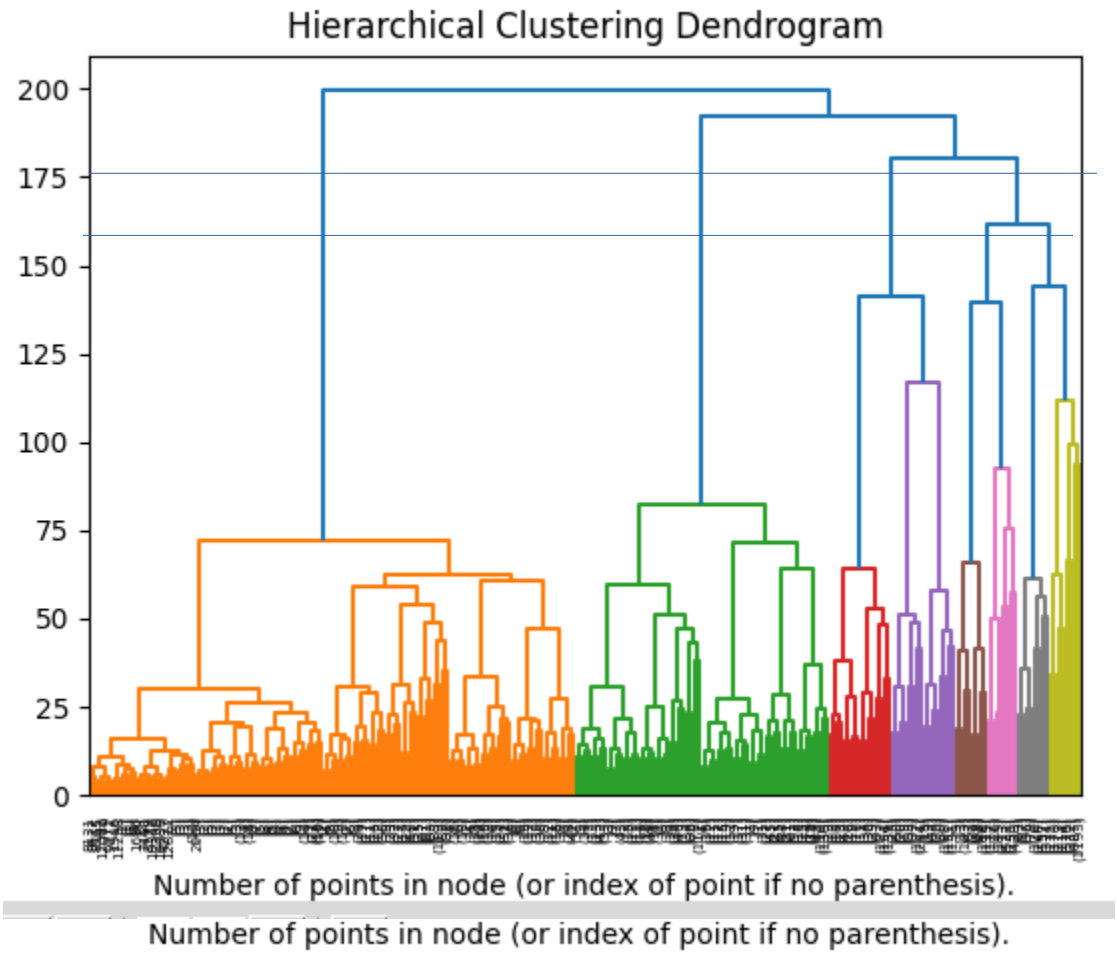
Изображена диаграмма "Каменная сыпь" в обратном порядке показывающая уменьшение числа кластеров с увеличением расстояния.



Последнее точка справа это слияние в 1 кластер. Большие промежуки это расстояния слияния кластеров. Самое большое расстрояние между 2 кластеров, затер для 5 и 7, количество объектов в них нам пока не известно.

Количество кластеров, номер слияния, id слеяния для кластера ветви, количество объектов в кластере, расстояние

13110	id_13108	id_13109	5825	d_141
13111	id_13103	id_13110	7214	d_151
9	13112 id_13094	id_13101	3010	d_157
7	13113 id_13111	id_13112	10224	d_178
5	13114 id_13096	id_13113	11056	d_188
3	13115 __leaf__	id_13114	11057	d_188
2	13116 id_13102	id_13115	13118	d_213



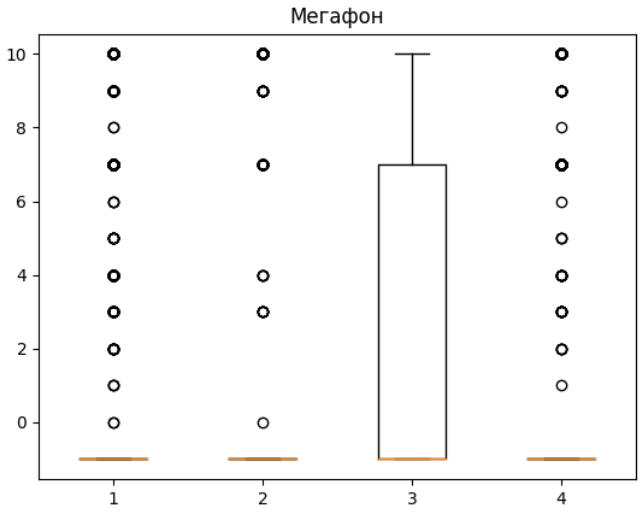
Number of points in node (or index of point if no parenthesis).

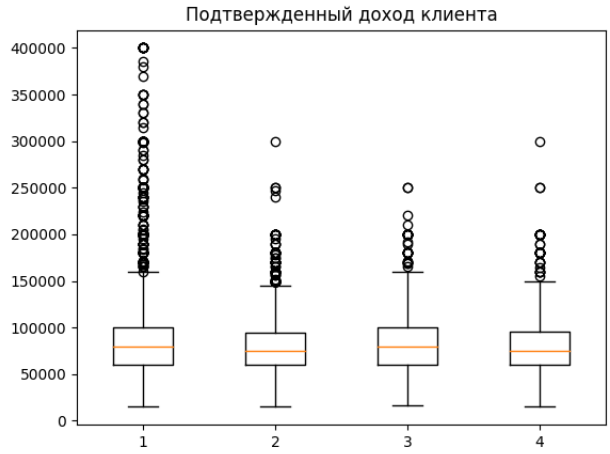
Первая диаграмма показыват нам излом где разряженность увеличивается и где мы ищем кластеры. На Второй диаграмме хорошо видна разряжцццценность кластеров, их объемы и расстояния друг от друга, после и в момент скачка разряженности.

Если провести черту на уровне 175 пунктов, то можно условно выделить 4 кластера. Попробуем их интерпретировать:

1 6894

2 2309
3 1945
4 1809

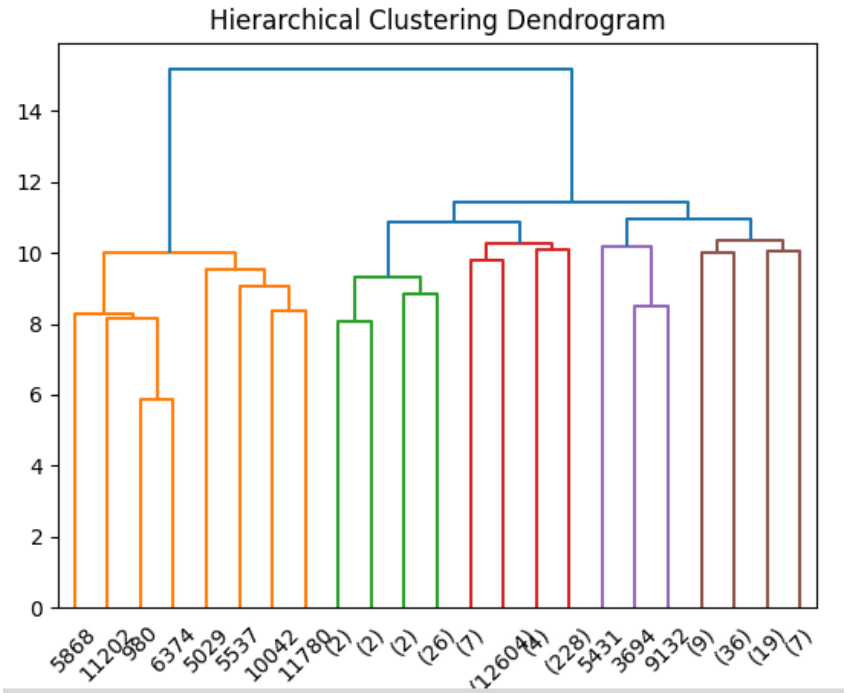


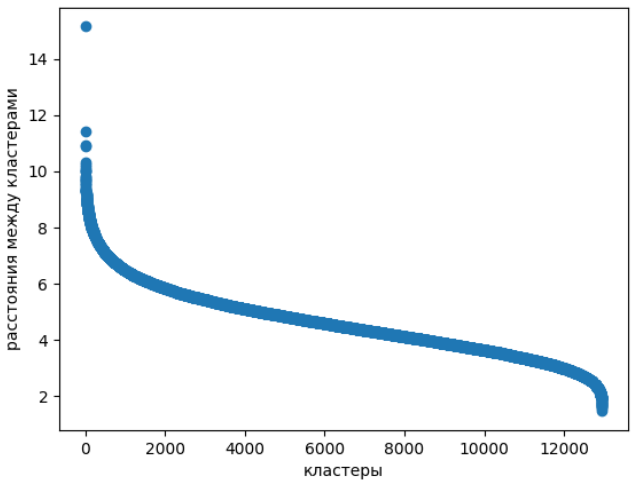


После сравнения квантилей столбцов и ящиковых диаграмм, значимых различий выявлено не было. Попытка интерпретировать не удалась, что говорит о случайном характере кластеров.

Поиск паровых скоплений

affinity='euclidean'
linkage='average'





Результат – один кластер который растет в размере (красный цвет).
Применим linkage='manhattan' – метрику устойчивую к выбросам. Результат тот же.

Поиск ленточных скоплений

affinity='euclidean' и 'manhattan'
linkage='single'
Результат неудовлетворительный – один кластер, который ростет в размере.

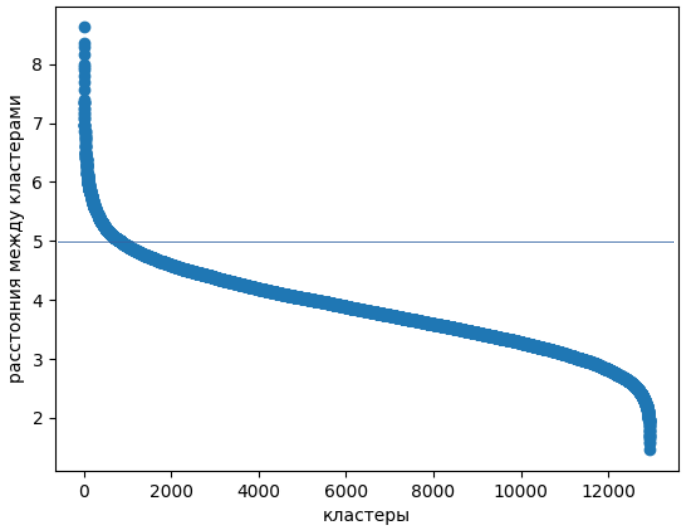
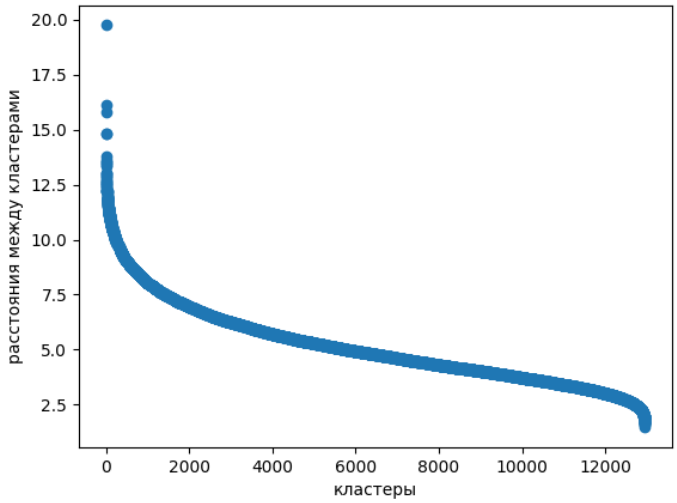


Диаграмма "каменистой сыпи" для метода ближайшей кластеризации. Рис 1

Методом поиска ленточный скоплений для объединения кластеров выбираются кластеры, которые имеют наиближайшую любую точку, это позволяет нам выделить группы точек которые имеют хотя бы одного близкого соседа. Такие точки можно интепретировать как заявки точно имеющие похожие рядом, а значит наиболее типичные. Заявки не имеющие соседей будем считат выбросами и отбросим. Возмем кластеры до излома на расстоянии 5, черта на рис. 1, и уже в этих заявках будем искать типичные значения параметров.

Поиск наиболее удаленных кластеров

affinity='euclidean' и 'manhattan'
linkage='complete'

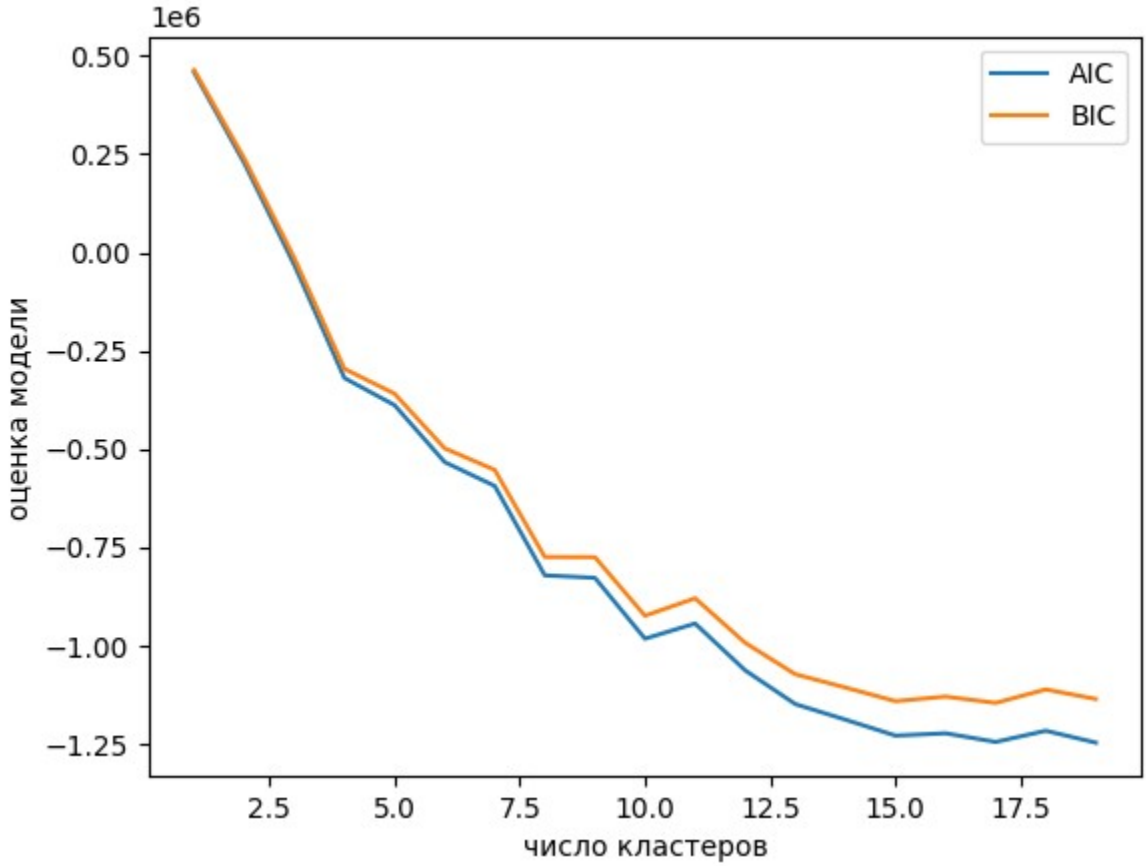


Снова один кластер, который медленно растет в размере .

Кластеризация EM-алгоритмом максимального ожидания.

Итерационный алгоритм апроксимации распределений переменных используя смеси гауссиан. Каждый кластер описывается центром, ковариациями (кластеры элиптической формы) и размером. Кластеры могут налагаться друг на друга.

Оценки AIC и BIC (их меньшие значения) позволяют сравнить модели и определить число кластеров. Однако для наших данных эти оценки монотонно убывают, не указывая число кластеров. Интерпретировать результат не удается, что говорит о неудаче кластеризации.



Метод распространения близости AffinityPropagation

Использующийся для нахождения числа кластеров, так же не дал результата. В серии экспериментов алгоритм всегда сходится через 100-200 итераций через 5 минут и находит 600-700 кластеров.

Итог кластерного анализа

Интерпритировать результаты кластеризации не удалось, так как методы показывают один растущий класер или несколько кластеров ничем друг от друга не отличающимися.

Методом локтя и ближайших соседей мы выбрали кластеры, которые эффективно исключают выбросы и позволяют оценить типичные одобренные андерраайтером заявки.

Из 12957 мы выбрали 12165 заявок. В результате мы получили такие данные, где 50% это медиана, показывающая типичную заявку, а 25% и 75% квартили показывающие медианы между тимичным значением и наименьшим и наибольшим значением соотственно.

Интеркварильный размах, дает нам возможность оценить границы, в которых андеррайтер, типично, принимает положительное решение.

	Запрошенная сумма кредита	Доход клиента	Подтвержденный доход клиента	Мегафон
count	12957	12957	12957	12957
mean	1335438.22566952	116034.976229065	82963.3909083893	0.790152041367601
std	478050.617502179	47480.6823900273	39676.5512209607	3.59833439952339
min	318078	35000	15000	-1
25%	981794	85000	60000	-1
50%	1293750	100000	80000	-1
75%	1617920	135000	100000	-1
max	3451480	600000	400000	10

	Скоринговый балл ОКБ, основной скоринг бюро	Эквифакс 4Score	Анкетный скоринг	Возраст клиента
count	12957	12957	12957	12957
mean	755.994211623061	732.521957243189	92.7284093540171	43.8587636026858
std	349.415337164051	111.462459952365	24.0875911280161	9.38629956023488
min	-168	0	35	23
25%	629	651	74	36
50%	823	734	90	43
75%	976	815	105	51
max	1837	952	183	67

	Месяц рождения клиента	Час создания заявки	Месяц создания заявки	подтв_минус_доход
count	12957	12957	12957	12957
mean	6.38697229296905	12.185073705333	6.20923053175889	33071.5853206761
std	3.4364485781285	3.16425643326214	3.68421798783368	34951.6233499367
min	1	5	1	0
25%	3	10	3	0
50%	6	12	6	25000
75%	9	14	10	50000
max	12	21	12	300000

	Скоринговый балл НБКИ общ

12957
551.122173342595
321.080304155437
0
552
688
781
850