

Определение доходности облигаций по данным ММВБ

Маричук Ю.В.

Август'22



Проект ПЕРЕЗАПУСК

О себе

- *Маричук Юлия Владимировна*
- *Московский Энергетический институт (Технический университет). Ф-т Автоматики и вычислительной техники.*
- *SberData, руководитель направления, реализация проверок качества данных в АХД (TeraData)*
- *К переезду не готова*
- *Marichuk.y.v@sberbank.ru*

Описание проекта

- *Необходимо оценить рынок рублевых облигаций по распределению купонной доходности, разделить их на кластеры и рассчитать доходность на период год по ограниченному скоупу облигаций (условия в задании).*
- *Ссылка на репозиторий с кодом*
https://github.com/YuliaMV/da_homeworks/blob/main/Project/Project_DA%20v4.ipynb

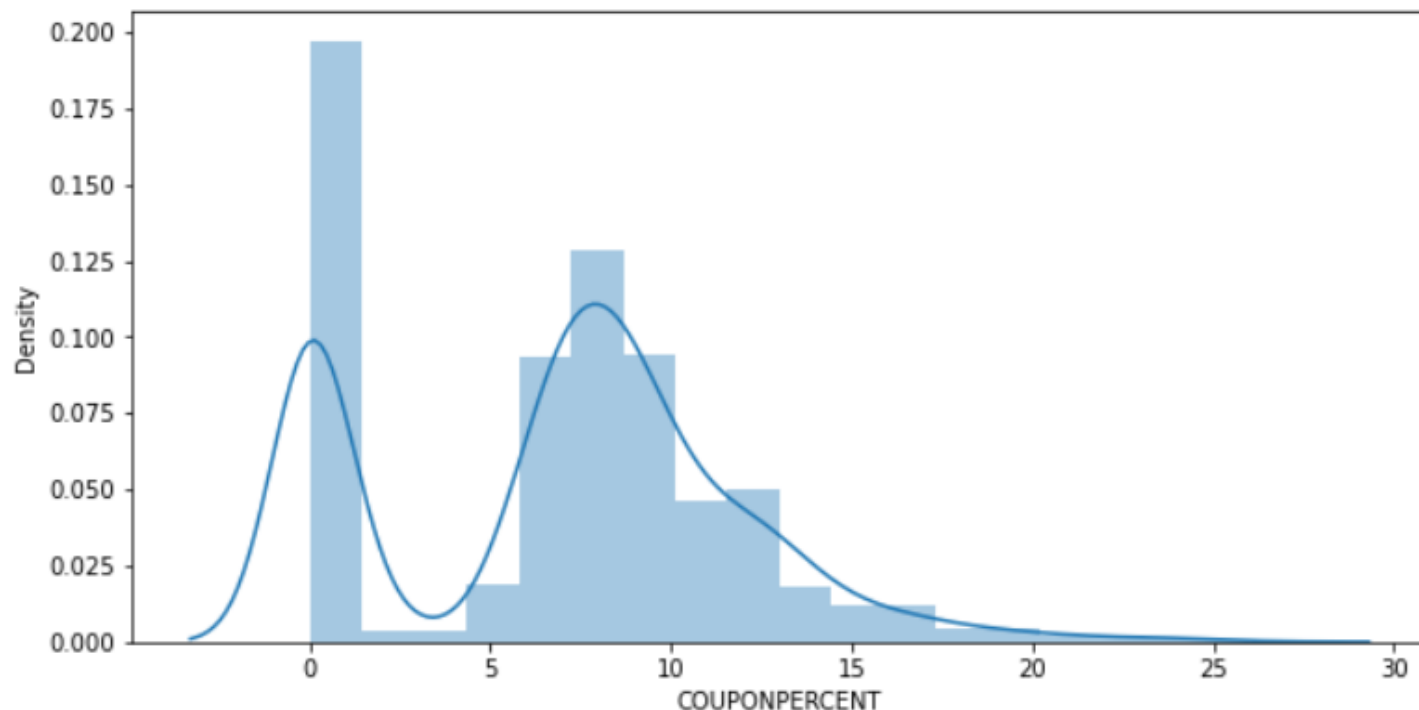
Бизнес-логика

- *Подготовка данных*
- *Анализ распределения купонной доходности с помощью статистик и визуализаций*
- *Выделение группы облигаций для более детального исследования*
- *Подбор оптимальных параметров, кластеризация датасета и анализ получившихся кластеров*
- *Расчет доходности с периодом год по облигациям, удовлетворяющим условию в задании*
- *Подбор оптимальных параметров, кластеризация датасета и анализ доходности получившихся кластеров*

Подготовка данных

- Оставили только рублевые облигации с заполненной ставкой купона
- Преобразовали дату в формат datetime
- Преобразовали числовые признаки в float
- Установили дату погашения, равной дате оферты, если она указана
- Посчитали дюрацию как срок до погашения, где не указано
- Оставили записи, где дюрация больше 0 и заполнено количество дней до погашения

Гистограмма и ядерная оценка плотности ставки купона

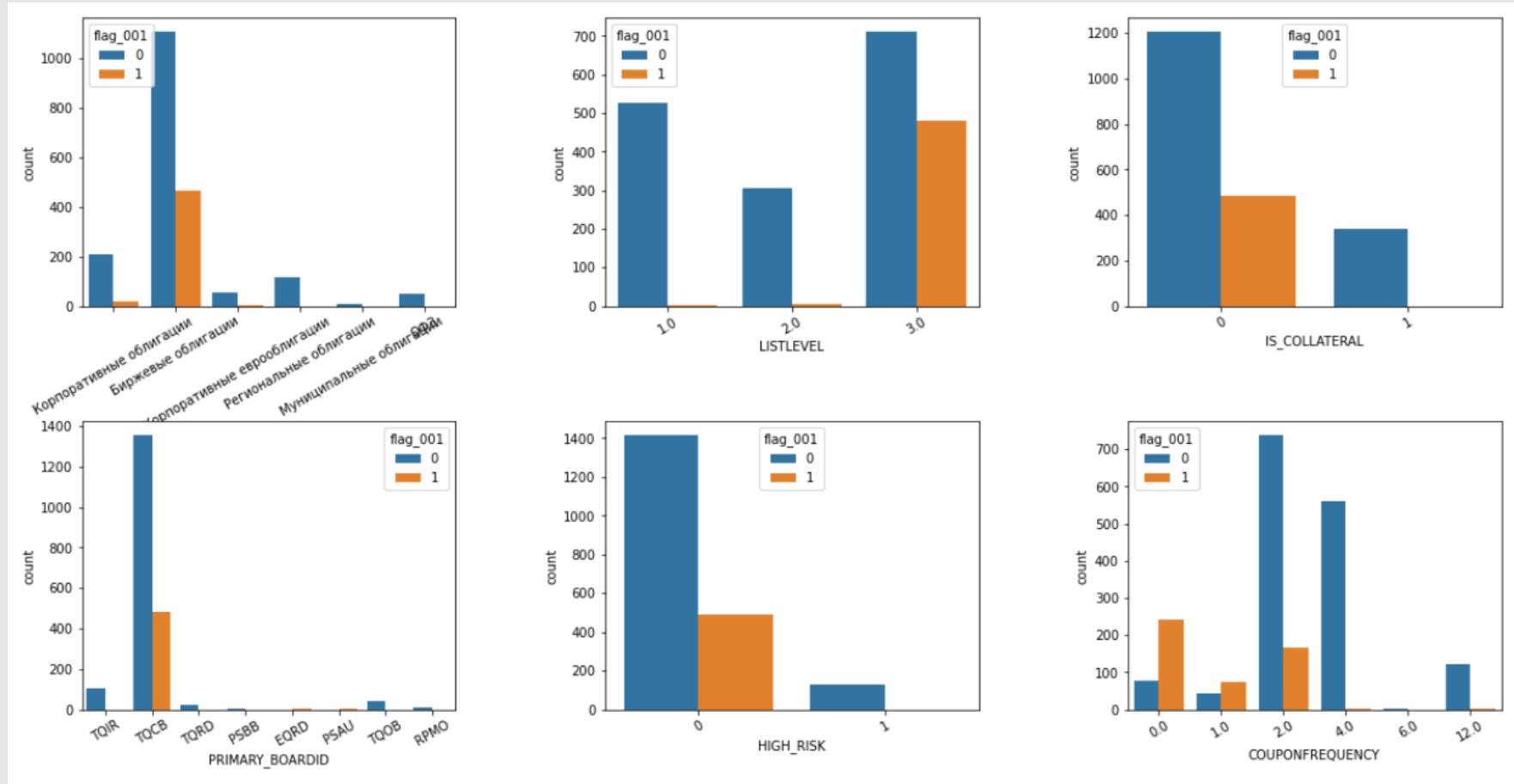


```
df.COUPONPERCENT.value_counts().head(10)
```

0.01	487
13.00	37
12.00	36
9.50	31
11.00	28
12.50	27
8.00	25
9.00	25
7.50	24
10.00	24

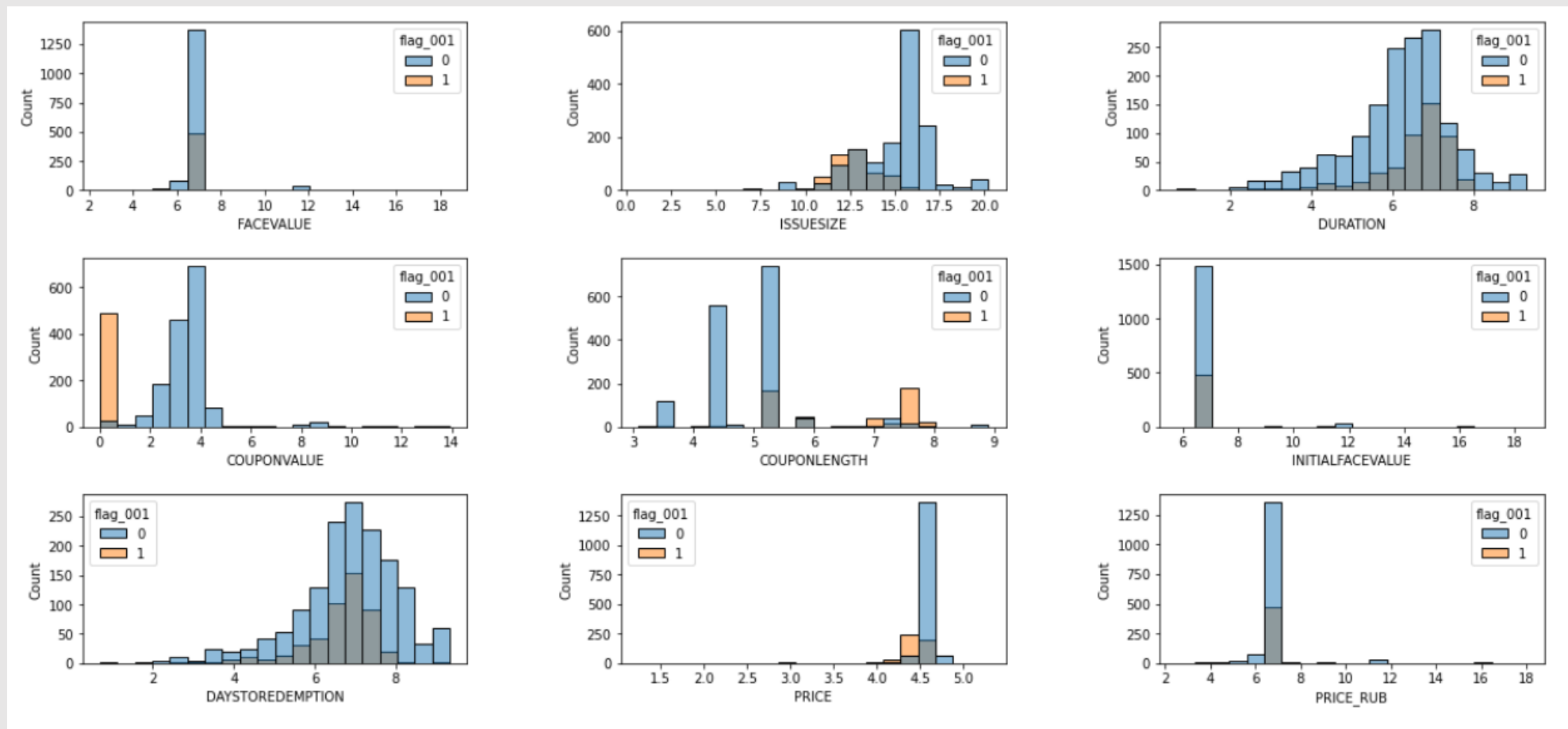
На графике присутствует выброс со значением 0.01 и два локальных максимума, возможно на распределение ставки влияют разные процессы. По цифрам видно, что облигации со ставкой купона 0,01 представляют отдельную группу. Для их выделения введем дополнительный признак `flag_001`. Посмотрим, как влияют числовые и категориальные признаки на этот флаг.

Зависимость признака flag_001 от категориальных признаков



Облигации со ставкой купона 0.01% низкорисковые (HIGH_RISK=0), только в 3-ем уровне листинга и без обеспечения (LISTLEVEL=3, IS_COLLATERAL=0)

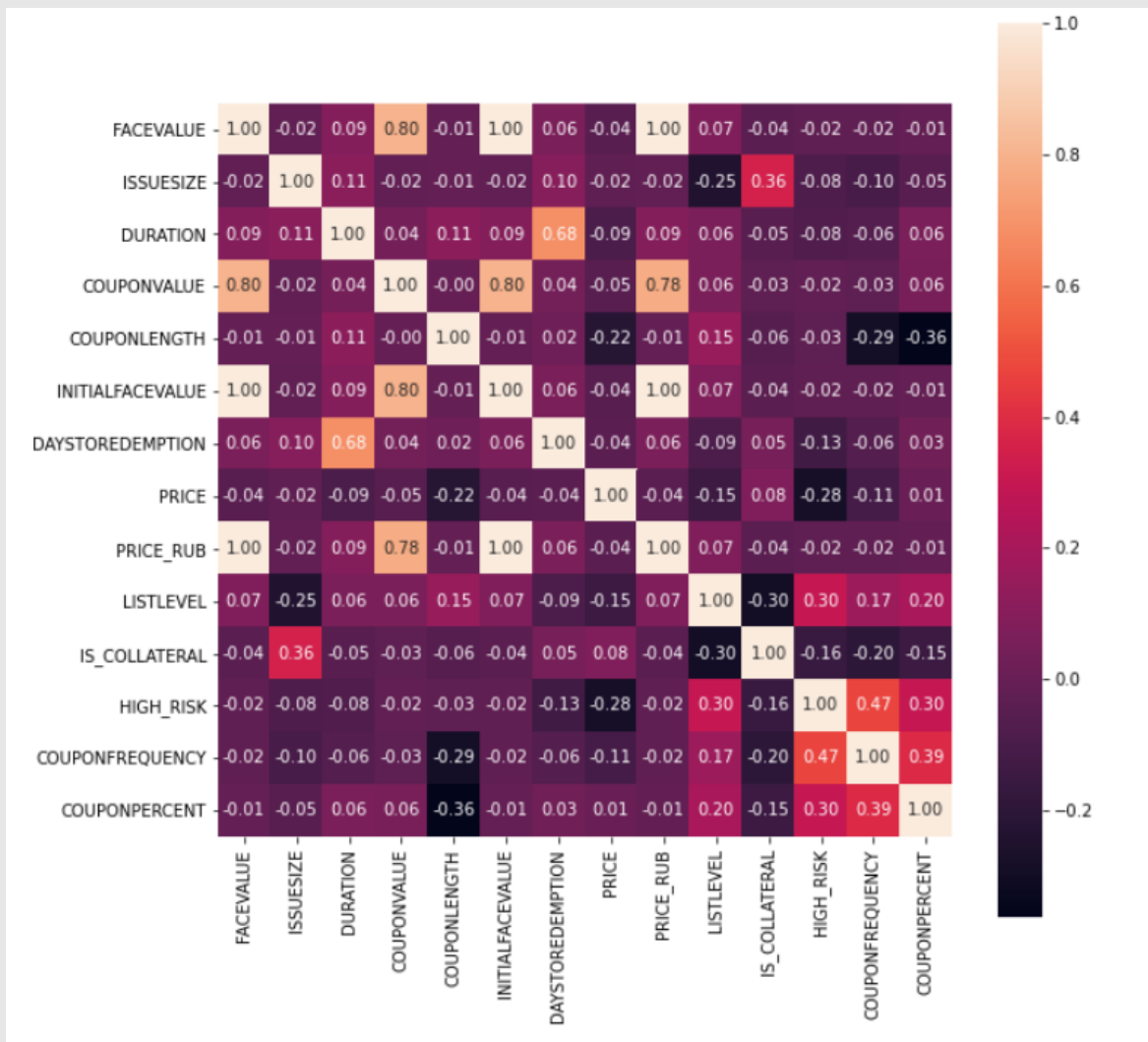
Гистограммы числовых признаков с группировкой по флагу flag_001



У облигаций со ставкой купона 0.01% всегда маленькая сумма купона (COUPONVALUE), также длительный купонный период (COUPONLENGTH) наблюдается только у этих облигаций.

В целом не удалось однозначно определить признаки, характерные только для облигаций со ставкой 0.01. Это отдельная группа облигаций, далее их исключим из исследования.

Матрица корреляций для flag_001 = 0



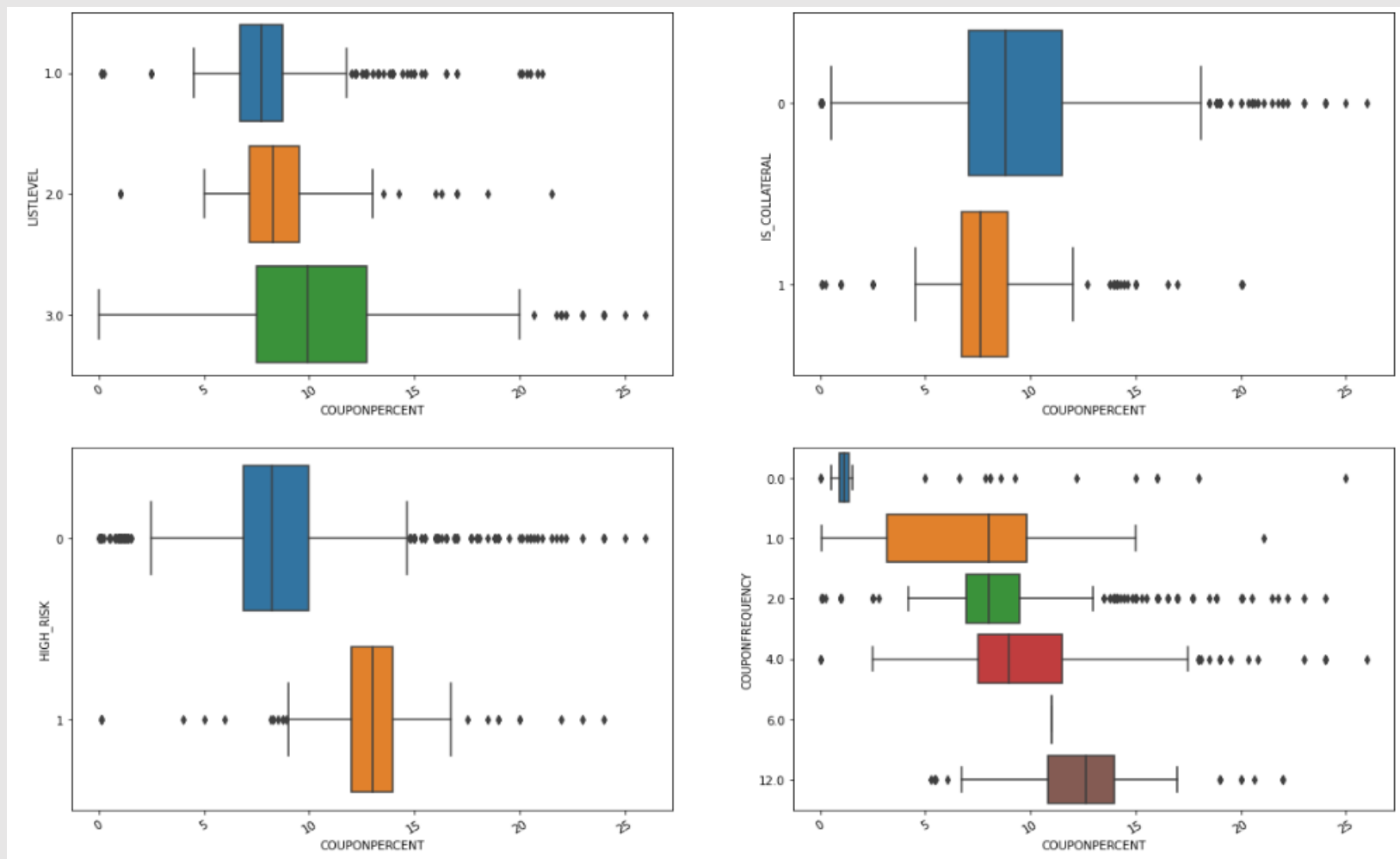
Положительная корреляция купонного дохода (COUPONPERCENT) с:

- частота выплаты купона (COUPONFREQUENCY)
- уровень риска (HIGH_RISK)
- уровень листинга (LISTLEVEL)

Отрицательная корреляция с:

- длина купонного периода (COUPONLENGTH)
- наличие обеспечения (IS_COLLATERAL)

Распределение на boxplot в зависимости от категориальных переменных



Некоторые группы имеют очень плотное распределение, почти везде медиана отличается от среднего. Сохранилось много выбросов.

Разделение облигаций на кластеры с помощью метода k-Means

- Признаки PRICE, DURATION, COUPONPERCENT
- Гиперпараметр n_clusters = 4

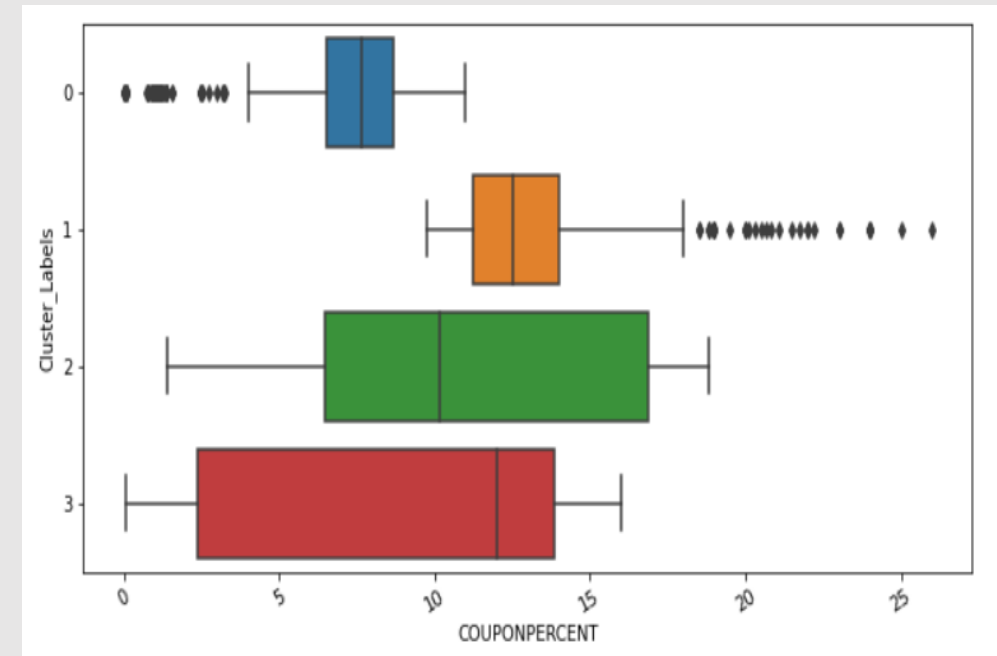
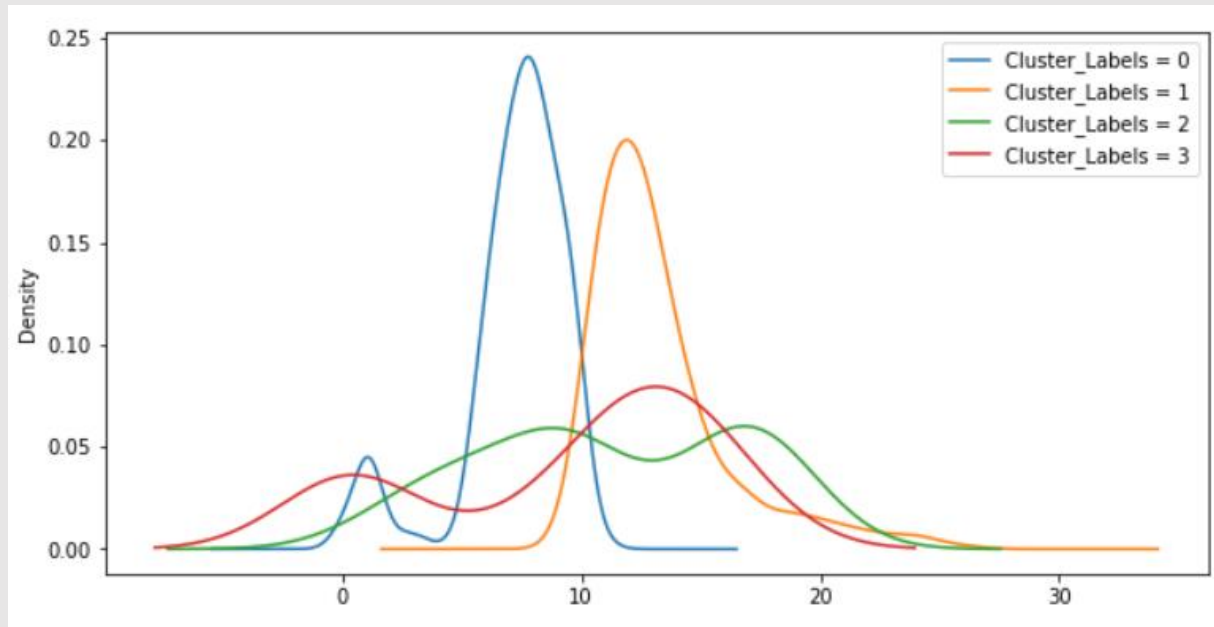
	count	mean	std	min	25%	50%	75%	max
Cluster_Labels								
0	1061.0	7.225433	2.268766	0.016	6.5500	7.65	8.700	11.00
1	403.0	13.276660	2.924519	9.750	11.2500	12.50	14.000	26.00
2	41.0	11.271220	5.392743	1.400	6.5000	10.15	16.900	18.83
3	34.0	9.598824	5.926427	0.100	2.3825	12.00	13.875	16.00

Нулевой кластер самый многочисленный, в нем облигации с низкими ставками купона, среднее и медиана ниже, чем в других кластерах.

Первый кластер второй по величине, в нем облигации с высокими ставками купона, среднее и медиана выше, чем в других кластерах. Дисперсия у этих кластеров не велика, распределение достаточно плотное.

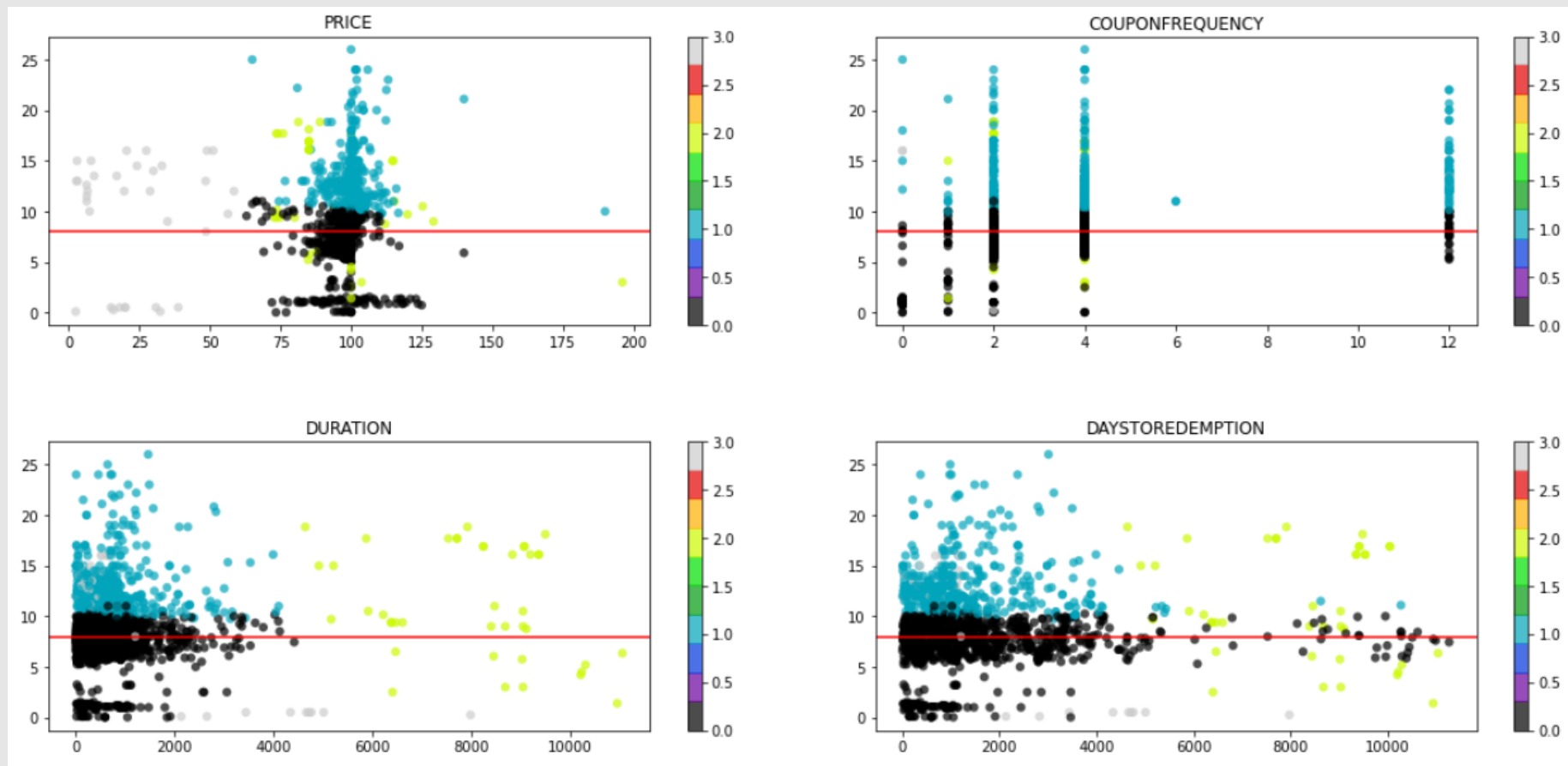
Два других кластера малочисленны и имеют большую дисперсию.

Графики распределения ставки купона и boxplot



Нулевой и первый кластеры имеют плотное распределение с выбросами, второй и третий "размазаны". Медиана второго больше смещена влево, медиана третьего вправо.

Зависимость ставки купона от дюрации, цены в %, количества дней до погашения и частоты выплаты купонов



0-й кластер (черный) имеет ставку около ключевой ставки ЦБ (красная линия) и ниже, не высокую дюрацию и цену в среднем ниже 100%

1-й кластер (голубой) имеет ставку выше ключевой ставки ЦБ и также не высокую дюрацию

2-й кластер (желтый) малочисленный, имеет большой разброс ставок, облигаций с ежемесячной выплатой купона в нем нет. Это долгосрочные облигации с высокой дюрацией и большим количеством дней до погашения

3-й кластер (серый) имеет низкую цену в %, облигации этого кластера со ставкой выше ставки ЦБ имеют маленькую дюрацию, со ставкой ниже – высокую. Возможно этот кластер можно еще разделить на два.

Расчет доходности

Формула доходности

$$Profit = \frac{((\text{Цена_номинальная} - \text{Цена_покупки}) * (1 \text{ ЕСЛИ } \leq 0 \mid 0.87 \text{ ЕСЛИ } > 0) - \text{НКД} + 0.87 * \text{Купонный_доход_руб} * X_{\text{лет}}) * 100}{((\text{Цена_покупки} + \text{НКД}) * (1 + \text{Комиссия_брокера_в_процентах}/100 + \text{Комиссия_биржи}/100))}$$

Статистические характеристики полученной величины:

- Очень большая дисперсия
- Есть отрицательная доходность
- Есть очень большая доходность

Выборочное среднее = 12.1736
Дисперсия = 772.1726
Стандартное отклонение = 27.7880
Минимум = -47.3128
Максимум = 462.3648
Медиана = 7.6781
Мода = 0.0085

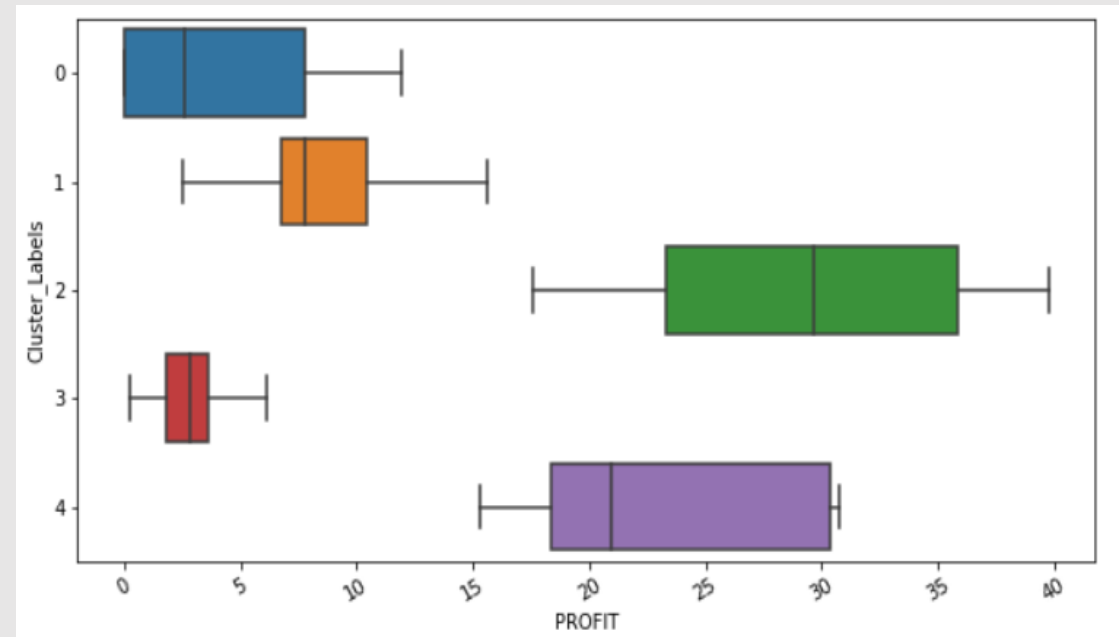
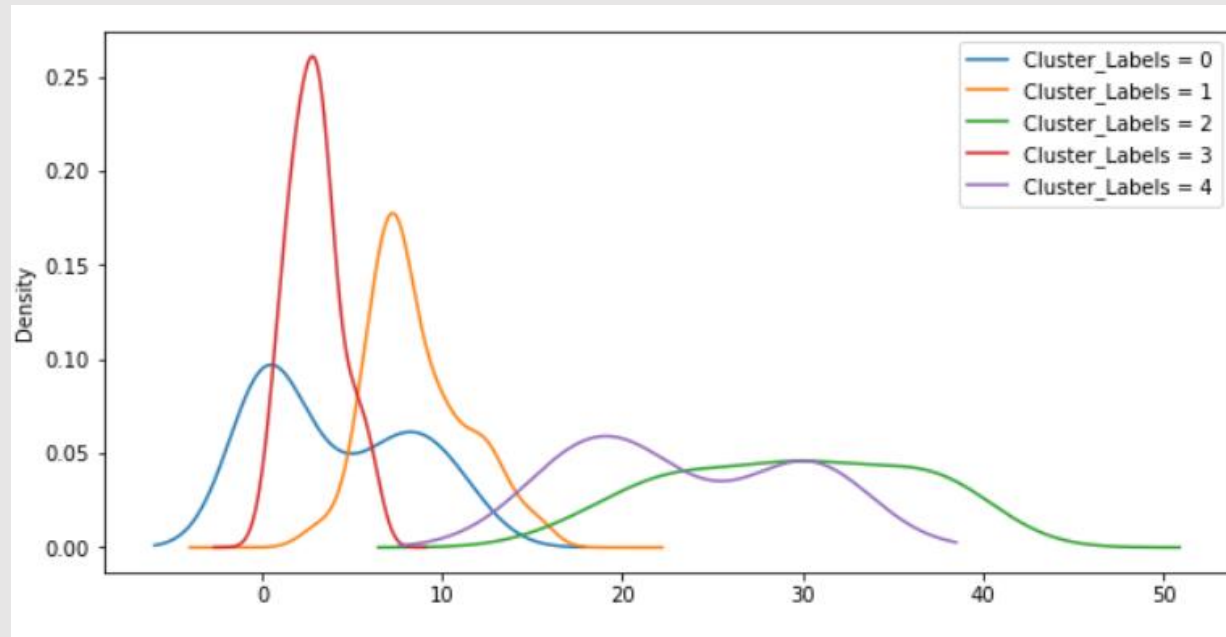
Далее не будем рассматривать облигации с доходностью меньше 0 и больше 40.

Кластеризация облигаций с помощью метода k-Means и распределение доходности

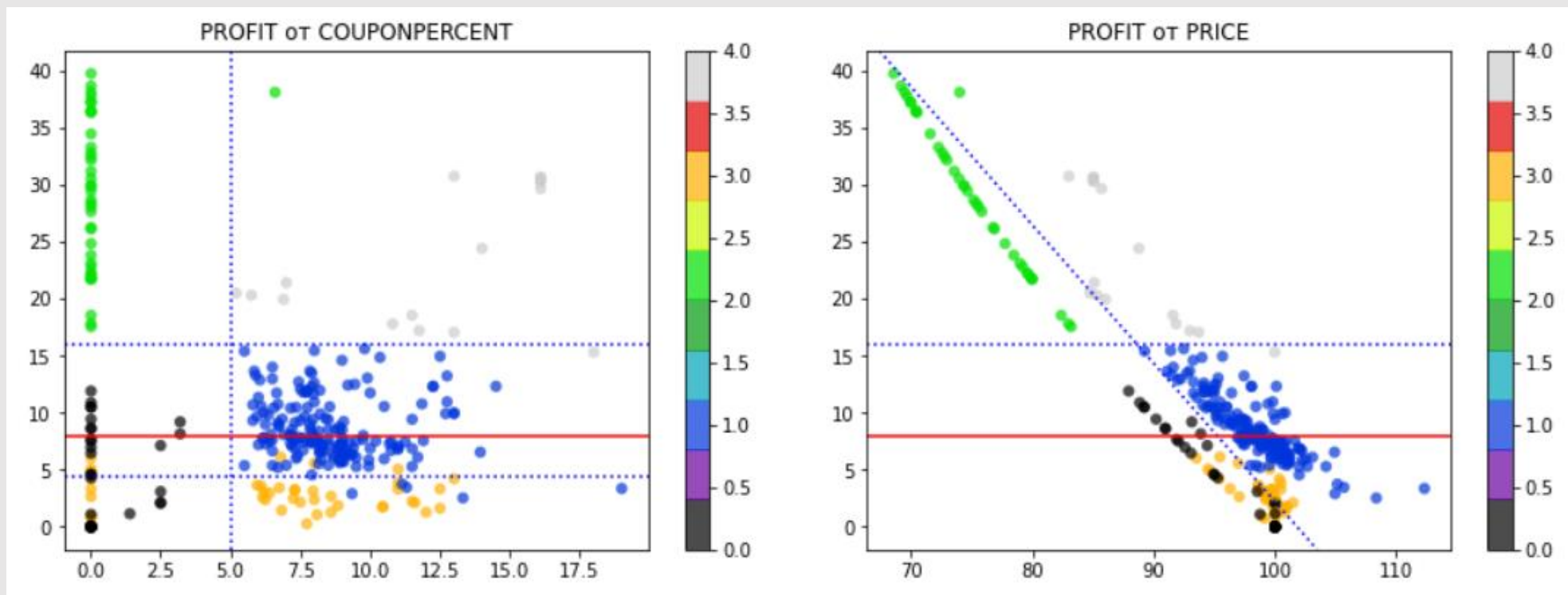
- Признаки PRICE, COUPONPERCENT, PROFIT, DAYS_TO_MATDATE
- Гиперпараметр n_clusters = 4

У кластеров 1 и 3 плотное распределение с небольшой дисперсией. Кластера 2 и 4 имеют большую дисперсию и более плоское распределение. Доходность в них высокая, не пересекается с доходностью в других кластерах, даже в хвостах

	count	mean	std	min	25%	50%	75%	max
Cluster_Labels								
0	36.0	4.081484	4.180802	0.000816	0.008022	2.625633	7.802378	11.919662
1	176.0	8.572907	2.686294	2.537208	6.784533	7.775246	10.430601	15.638851
2	38.0	29.387916	6.569362	17.564409	23.313154	29.699392	35.901595	39.796517
3	40.0	2.930402	1.442007	0.277429	1.856311	2.829479	3.599282	6.135566
4	16.0	23.454193	5.938708	15.328116	18.383909	20.980474	30.352084	30.784134



Зависимость доходности от ставки купона и от цены в %



- Нулевой кластер (черный) имеет низкие ставки купонов, цену ниже номинала и в среднем доходность ниже ключевой ставки ЦБ (красная линия).
- Первый кластер (синий) самый многочисленный имеет доходность вокруг ключевой ставки ЦБ от 5% до 16%
- Второй кластер (зеленый) имеет маленькую купонную ставку, цену сильно ниже номинала и большую доходность. Доходность высокая возможно за счет большой разницы между ценой номинала и ценой покупки, а возможно за счет высокого НКД
- Третий кластер (желтый) имеет доходность ниже ключевой ставки ЦБ
- Четвертый кластер (серый) имеет высокую купонную ставку, цену ниже номинала и высокую доходность

Доходность облигаций во 2-ом и 4-ом кластерах сильно выше ключевой ставки ЦБ

Используемые технологии

Проект реализован на Python в Jupyter-ноутбук. Использовались библиотеки Python:

- pandas, numpy - для расчетов и подготовки данных*
- matplotlib, seaborn - для визуализаций*
- кластеризация проведена методом k-Means*

СПАСИБО ЗА ВНИМАНИЕ!