

# ПРЕДСКАЗАНИЕ ВИДИМОСТИ МУЗЕЙНОГО ОБЪЕКТА

НА ОСНОВАНИИ ДАТАСЕТА ART INSTITUTE OF CHICAGO API



# КОМАНДА ПРОЕКТА



**Василиса Попова**

Тимлид, ответственная за временные признаки



**Анна Лебедева**

Ответственная за признаки активности работы



**Елизавета Доценко**

Дизайнер, ответственная за художественные признаки



**Кирилл Орлов**

Ответственный за биографические признаки автора

**American Gothic, 1930**

Grant Wood

Источник: АИС



# OVERVIEW

01 EDA

02 BASELINE

03 ОБРАБОТКА АНОИМАЛИЙ. FEATURE  
ENGINEERING

04 ОТБОР ПРИЗНАКОВ

05 МОДЕЛЬ ЭТАПА 2

06 ИНТЕРПРЕТАЦИЯ ФИНАЛЬНОЙ МОДЕЛИ

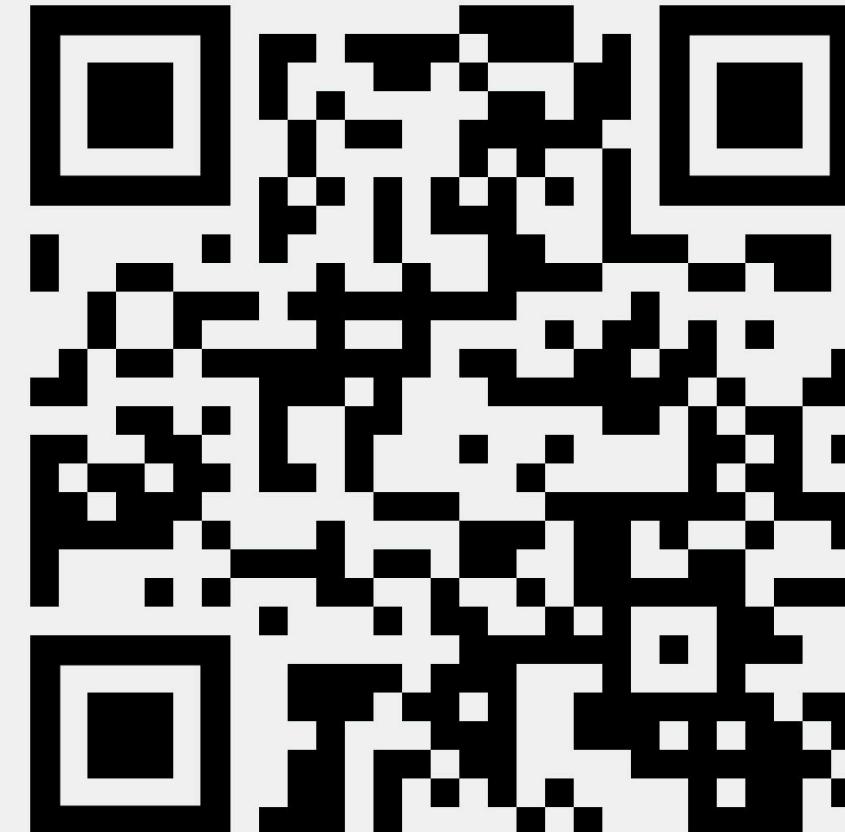


**Hero Construction, 1958**

Richard Hunt

Источник: AIC

# СБОРКА ДАТАСЕТА



## Источник

API Art University of Chicago

## Преобразование

Распарсили JSON по работам и авторам, агрегировали по каждой работе признаков видимости (выставки/публикации/продукты), добавили географию и время и сохранили как финальный CSV

## Выбор таргета

Таргет visible задаётся как 1, если у работы выполнено хотя бы одно из условий: was\_exhibited > 0 или in\_product > 0 или in\_publication > 0, иначе 0.

# Видимость состоит из

Бизнес-цепочка музея

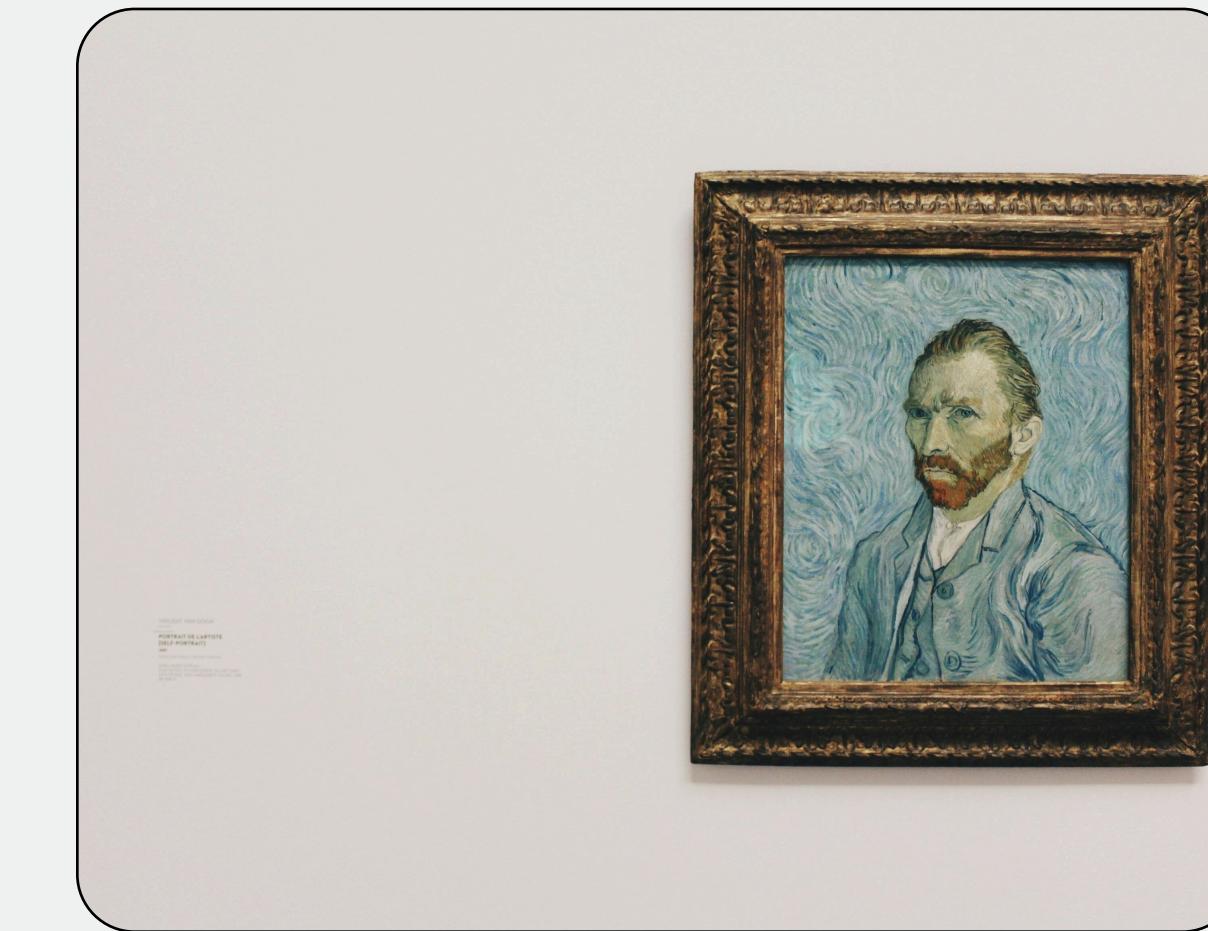
↓  
М1-модель видимости

↓  
Кураторские/  
продуктовые  
решения

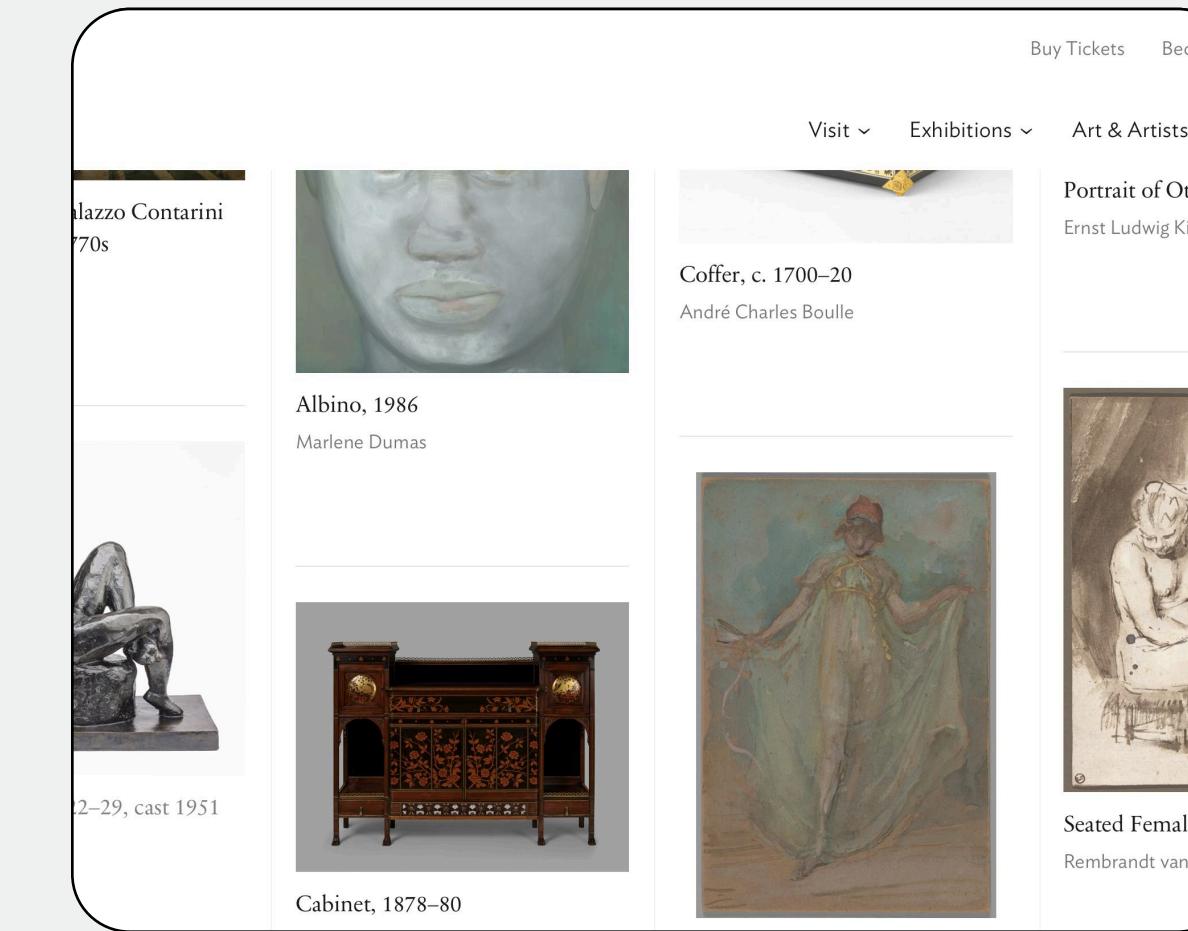
↓  
Подготовка

↓  
Каналы

↓  
Видимость/  
доход/  
вовлеченность



Выставочной  
деятельности



Публикационной  
деятельности



Массовой  
видимости

**Ценность для бизнеса**

Мы не пытаемся заменить кураторскую экспертизу.

Мы находим объекты, которые по наблюдаемым признакам статистически похожи на те, которые в прошлом становились видимыми, ещё до того, как музей вложит ресурсы в выставку, публикацию или продукт.

# EDA

---

ARTSCOPE

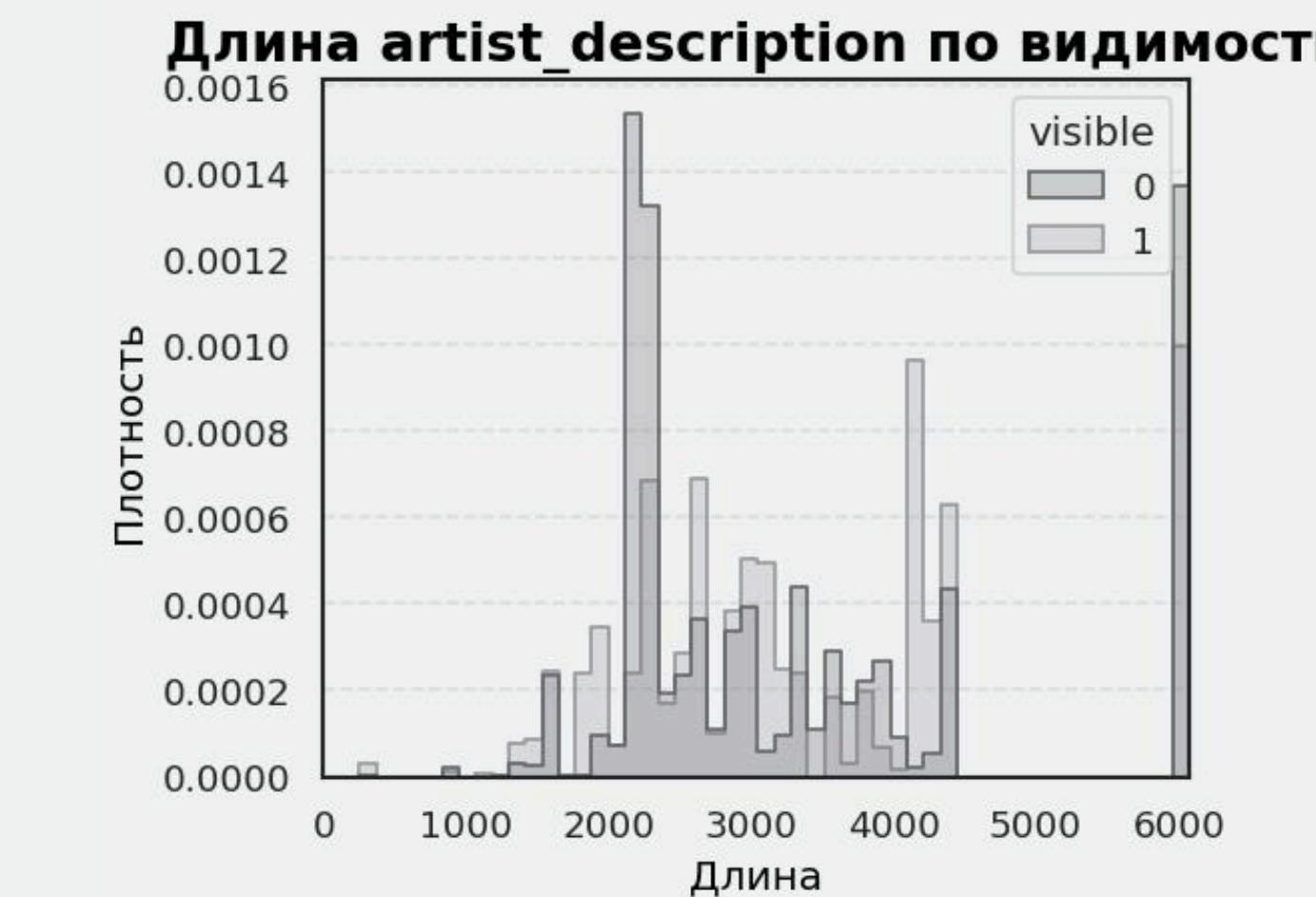
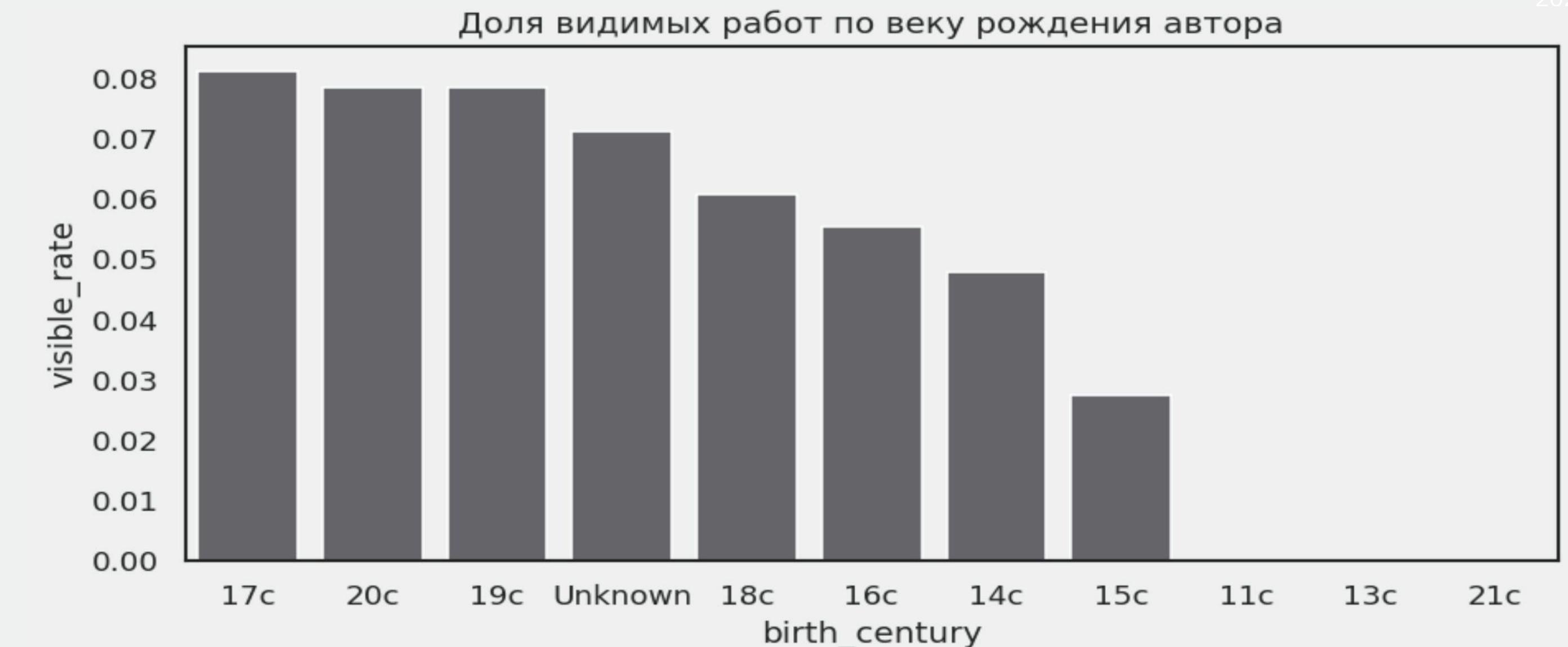


Кирилл Орлов

## АВТОРСИКИЕ ПРИЗНАКИ

Много выбросов. Ключевой признак - происхождение автора. В целом данные шумные, нужно хорошо подготовить.

Feature	% NaN
agent_type_id	100%
artist_gender	100%
artist_nationality	100%
artist_description	90%



ARTSCOPE

2021



Елизавета Доценко

# ХУДОЖЕСТВЕННЫЕ ПРИЗНАКИ

Стили, департаменты, классификация - признаки с сильным сигналом.

## Classification title

1. Гравюра (11529)
2. Литография (10796)
- ...
6. Фото (6769)
- ...
13. Живопись (2072)

## Style title

1. Японский (7864)
2. 21-ый век (4549)
- ...
6. Модернизм (1518)
- ...
9. Поп-арт (610)

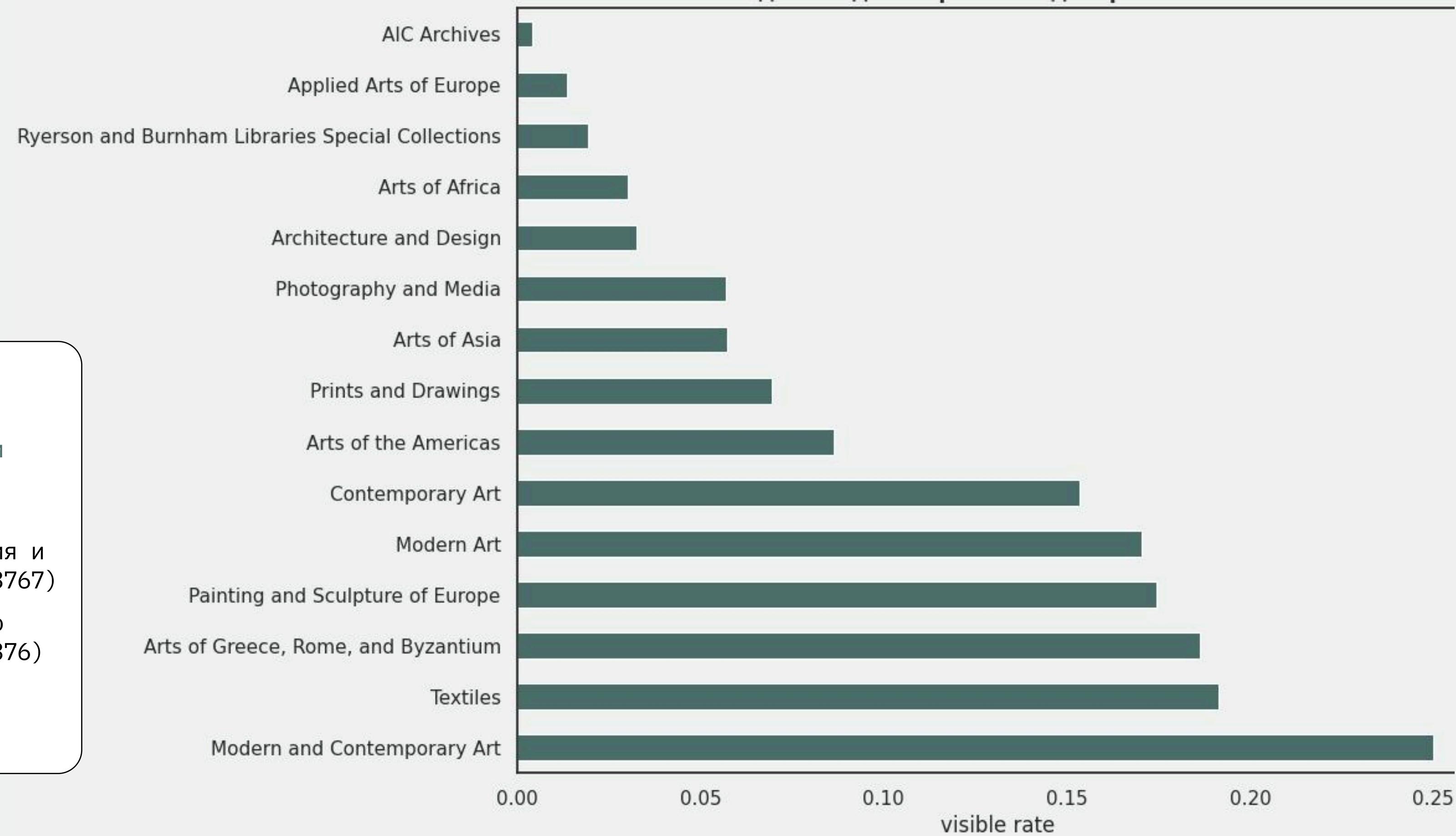
## Department title

1. Гравюры и рисунки (52164)
2. Фотография и медиа (23767)
3. Искусство Азии (16376)
4. Текстиль (11554)

Всего: 688  
Топ-20: 66%  
NaN: 1678

Всего: 625  
Топ-20: 93%  
NaN: 95803

Всего: 16  
NaN: 6568



ARTSCOPE



Елизавета Доценко

## ХУДОЖЕСТВЕННЫЕ ПРИЗНАКИ

Feature engineering

`dimension_text`. Из текста вида ‘*Each image: 9.2*

- ✓  $\times 7.4 \text{ cm } (3 \frac{5}{8} \times 2 \frac{15}{16} \text{ in.})\dots$ ’ в `normalised_side` - нормированную сторону в метрах

- ✓ `thumbnail_area`. Из ширины и длины превью в его площадь.

- ✓ Для описания картины и названия классификации – Multi-hot encoding по топ-40 словам.

- ✓ `title`. Количество слов в `title_word_count`, vectorizer (Tf-DF + SVD)

- ✓ `credit_line` – количество слов



**Coverlet, 1843**

James Cunningham

Источник: AIC

ARTSCOPE

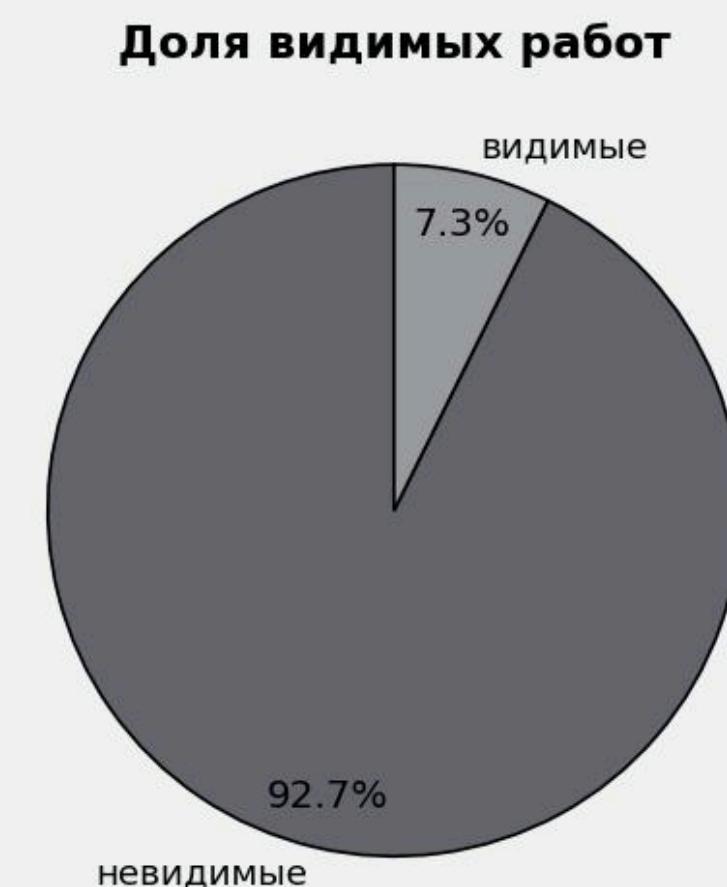


Анна Лебедева

## ПРИЗНАКИ АКТИВНОСТИ

**Сильная несбалансированность выборки.** Доля видимых работ составляет примерно 7%. Это означает, что большинство объектов никогда не было выставлено или задействовано в цифровых формах

Feature	% NaN
product_types	100%
time_since_last_exhibition	93.8%
last_exhibition_year	93.8%





Василиса Попова

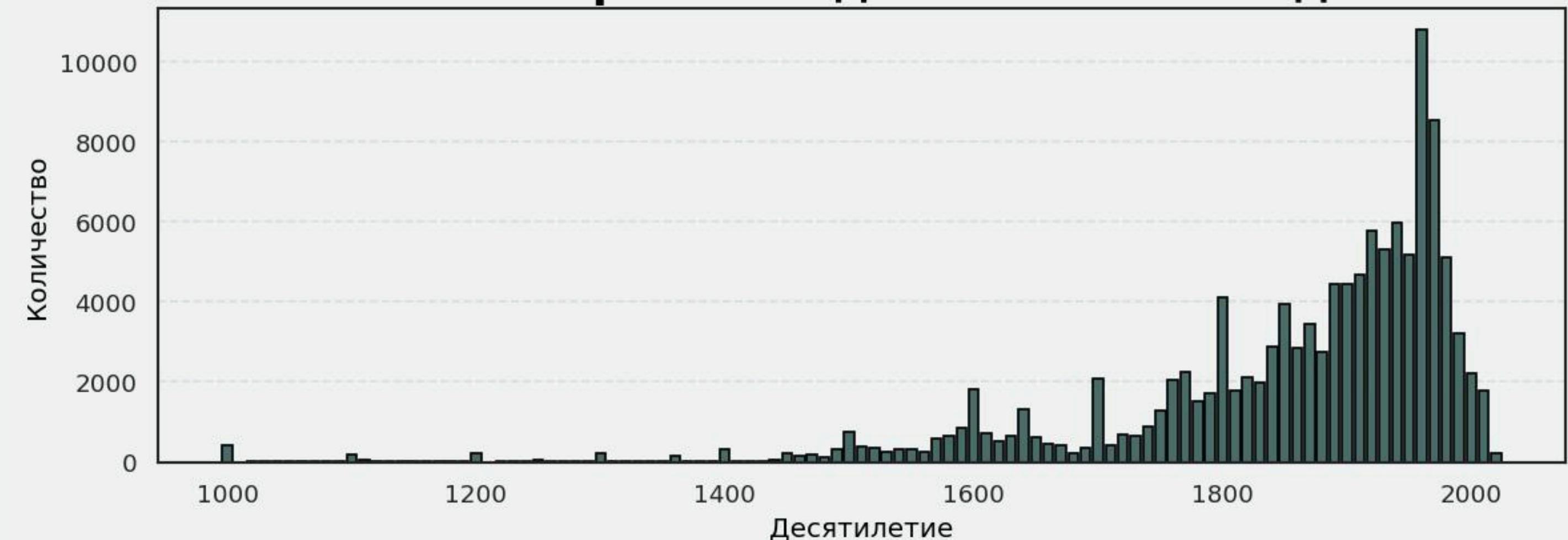
## ВРЕМЕННЫЕ ПРИЗНАКИ

Возраст работы связан с видимостью нелинейно: у молодых объектов она ниже, затем растёт и стабилизируется в диапазоне 50–200 лет, поэтому возраст стоит бинировать. Лаги и поля last\_\* хорошо описывают музейную историю, но в прогнозе дают прямую утечку.

## ГЕО-ПРИЗНАКИ

99.7% записей не имеют координат, а у 422 работ координаты музея (а не страны происхождения)

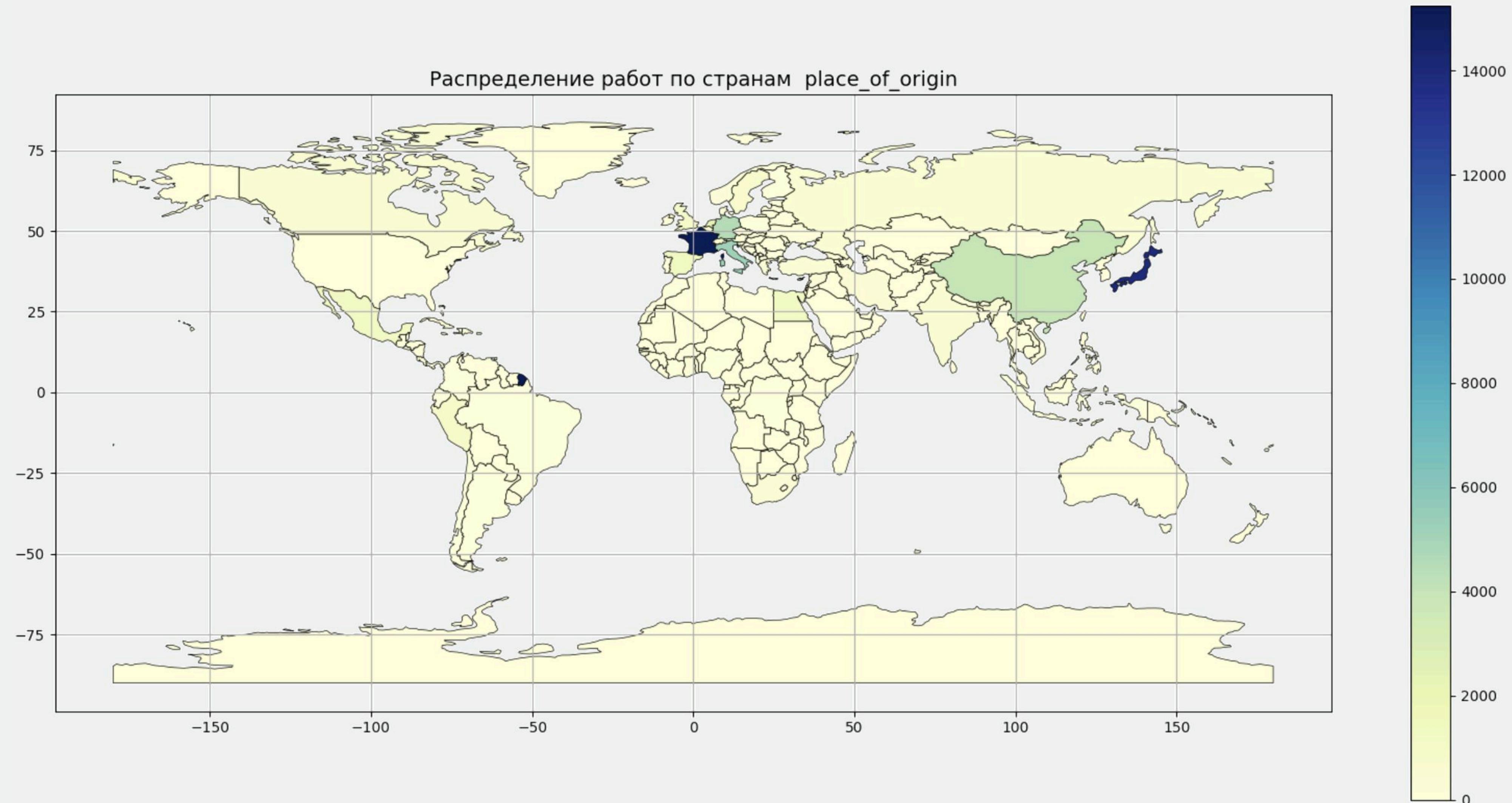
Количество работ по десятилетиям создания



Доля visible по десятилетиям создания



# ГЕО-ВИЗУАЛИЗАЦИЯ: КАРТА МИРА ПО PLACE\_OF\_ORIGIN



EDA показал, что видимость лучше всего объясняют три блока: авторский профиль, параметры объекта и историческое время.

# BASELINE.

## CatBoostClassifier

depth=6, lr=0.1, loss\_function="Logloss"

**0.93** ROC AUC

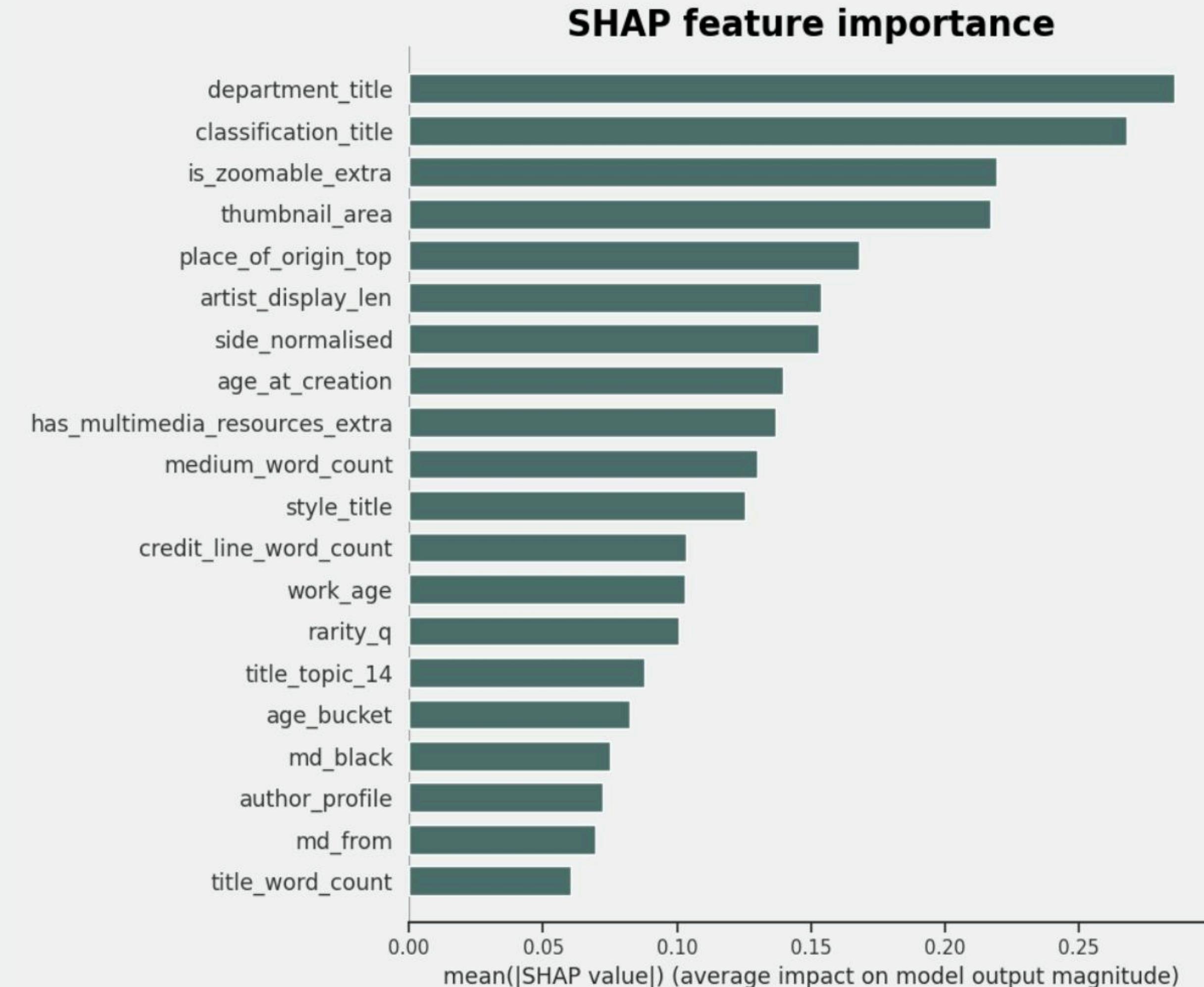
ст. отклонение на KFold = 0.0032

**0.60** F1

ст. отклонение на KFold = 0.0042

**134078** STRINGS

**84** FEATURES



# Обработка аномалий. Feature engineering



Two Ladies at the Automat  
(New York City), 1966  
Diane Arbus  
Источник: АИС

# FEATURE ENGINEERING

## IQR+Z-score

side\_normalised,  
thumbnail\_area,  
medium\_word\_count,  
title\_word\_count,  
**is\_outlier**

## Isolation forest

side\_normalised,  
thumbnail\_area,  
medium\_word\_count,  
title\_word\_count,  
**anomaly\_iforest\_flag**,  
**anomaly\_iforest\_score**

## Local Outlier

side\_normalised,  
thumbnail\_area,  
medium\_word\_count,  
title\_word\_count

## OHE

author\_profile,  
place\_of\_origin\_top,  
classification\_title\_top, style\_title\_top

## Target Encoding

place\_origin\_te,  
place\_origin\_te

## KNN-features

knn\_mean\_visible,  
knn\_rarity\_density,  
knn\_author\_cluster,  
artist\_rarity,  
birth\_century,  
desc\_len,  
knn\_visibility\_density,  
knn\_geo\_density

## Scaling

norm\_n\_documents,  
norm\_n\_sites,  
norm\_n\_texts,  
log\_exhibition\_count,  
log\_product\_count,  
log\_n\_documents,  
log\_n\_sites,  
log\_n\_texts

## Flags

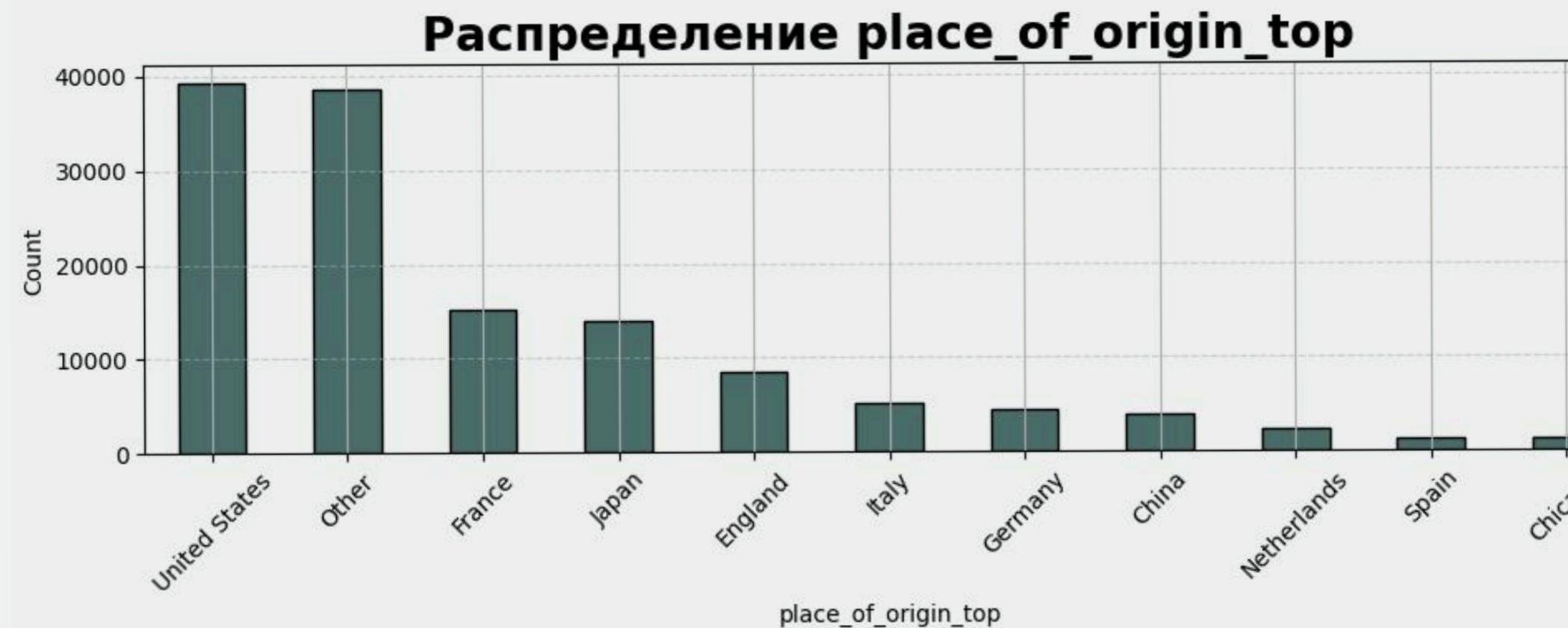
anomaly\_flag,  
age\_outlier,  
is\_unknown\_author,  
super\_exhibited,  
super\_products,  
super\_documents,  
super\_sites,  
any\_time\_negative\_lag

Что интересного  
мы заметили...

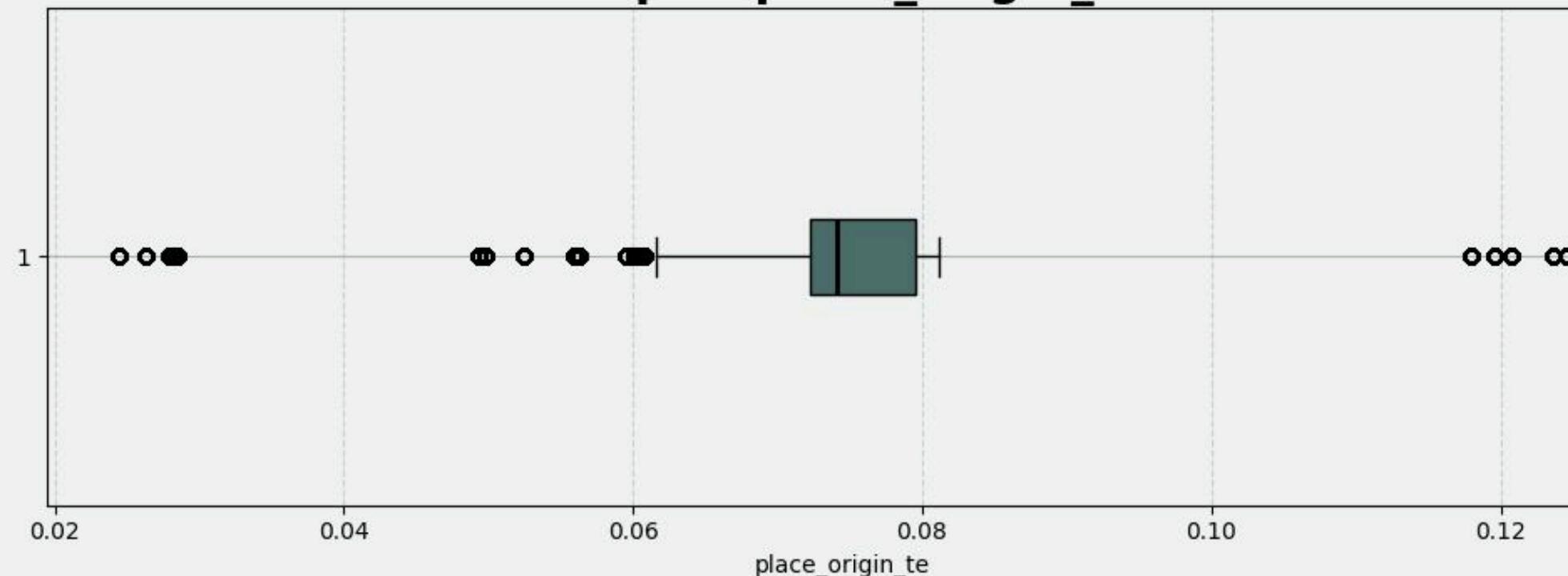
**Oil or Vinegar Cruet, 1737**  
Meissen Porcelain Manufactory  
Источник: АІС



## KFold Target Encoding места рождения

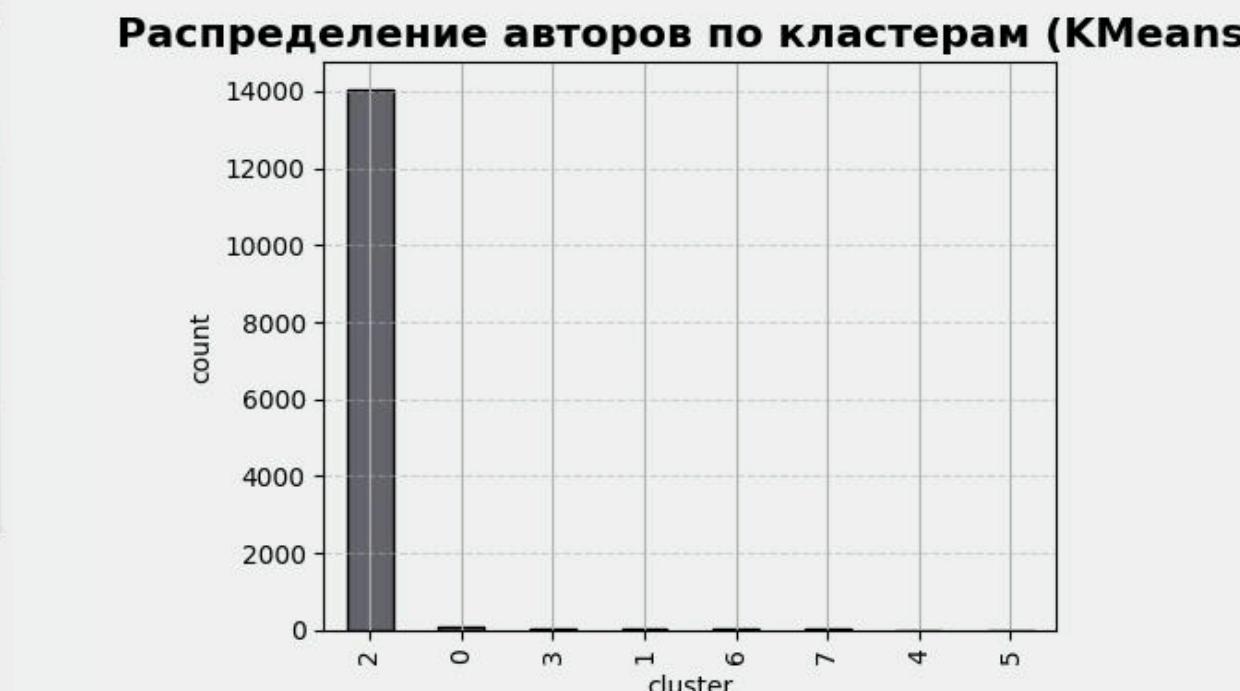


**Boxplot place\_origin\_te**



## TE автора

### KNN clusters для автора



## FE для профиля автора

author_profile	%	visible
dates_only	68%	7.43%
unknown_artist	17%	7.32%
no_info	15%	6.83%

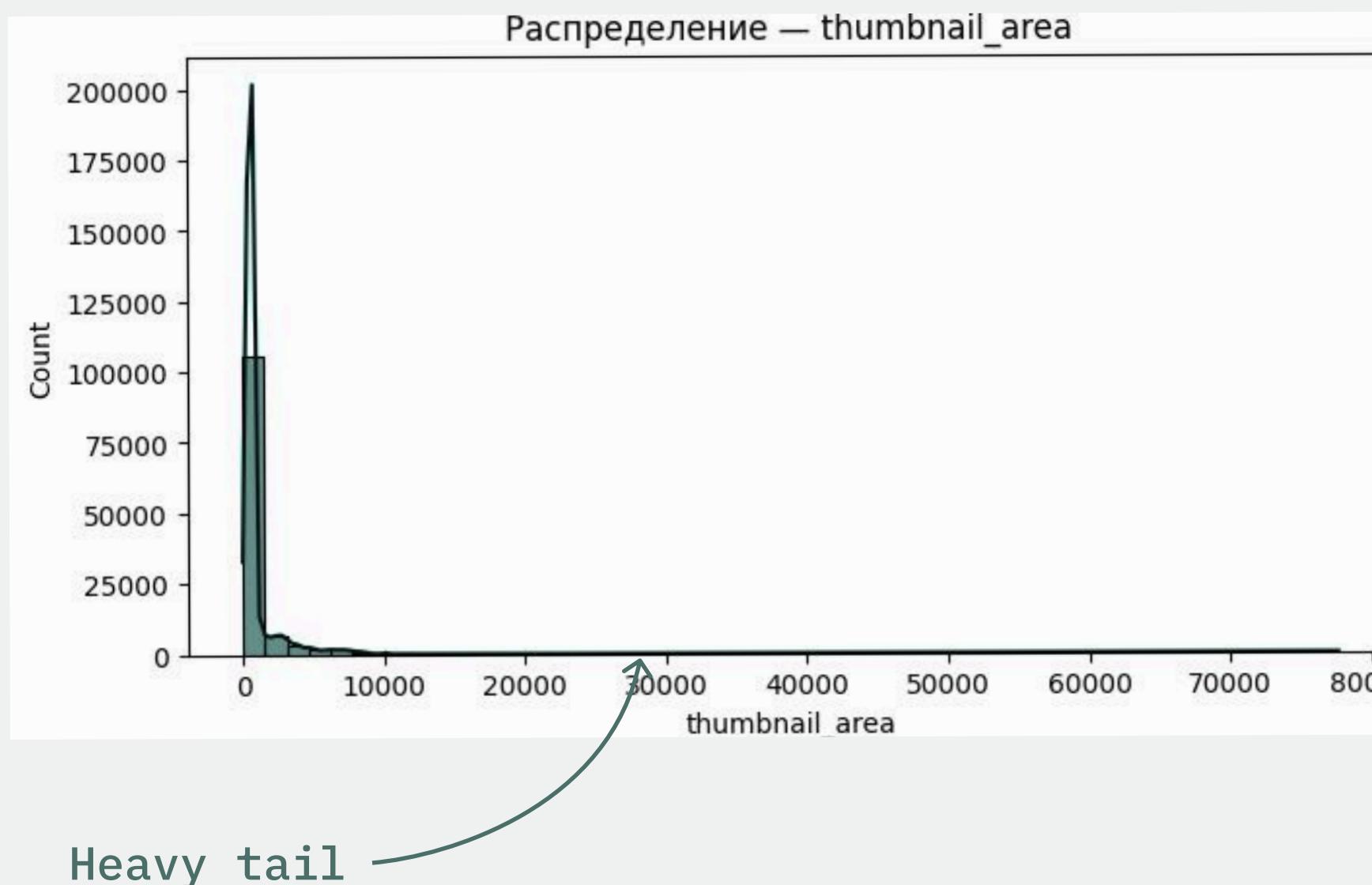
# ИТОГИ

по автору

Все добавленные  
биографические признаки и  
их корреляции

Feature	visible
place_origin_te	0.04047
knn_author_cluster	0.03622
knn_rarity_density	0.00858
is_unknown_author	-0.00006
author_profile_te	-0.00388
age_oulier	0.01344

## Выбросы по IQR в размере картины, площади миниатюры, длины описания и длины названия



Доля выбросов среди видимых объектов	55%
Доля выбросов среди не видимых объектов	29%

## TE для department

Топ по корреляции с таргетом до TE	Топ по корреляции с таргетом после TE
Photography and media	Textile
Textile	...
...	Photography and media

## TruncatedSVD + KMeans на medium\_display

Вспомним: во время EDA были созданы 40xmd\_\*.

На текущем этапе:

- После TruncatedSVD: 5xmd\_svd\_\*
- После KMeans: 1xmd\_claster - числовой признак, принимающий значения от 1 до 6.

# ИТОГИ

по художественным признакам

Все добавленные художественные признаки

Feature
is_outlier
anomaly_iforest_flag
anomaly_iforest_score
department_te
md_svd_*
md_cluster

## Топ категорий активности по пропускам

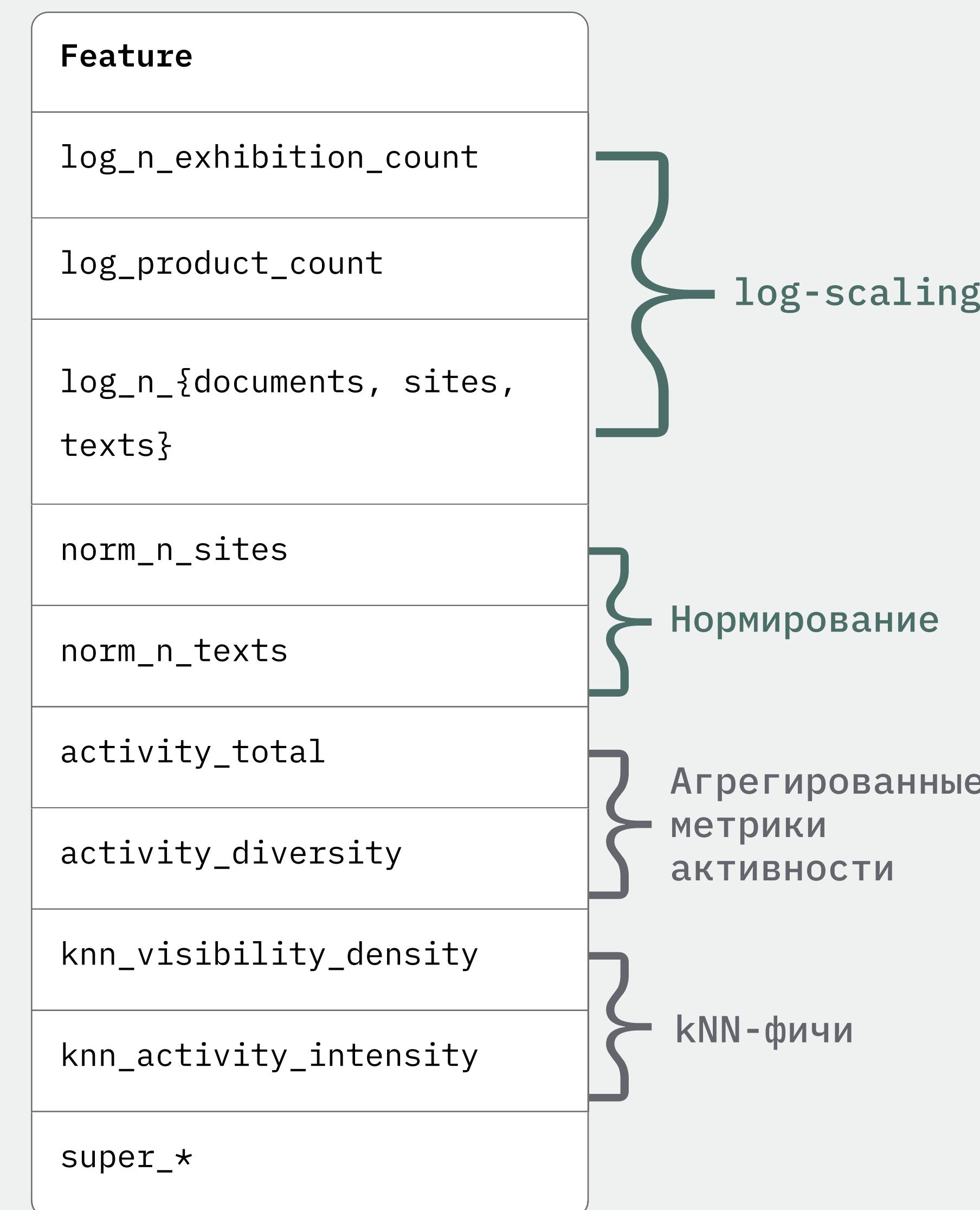
Feature	% NaN
product_types	100%
last_exhibition_year	93.8%
time_since_last_exhibition	93.8%



### Leak

Beth Van Hoesen

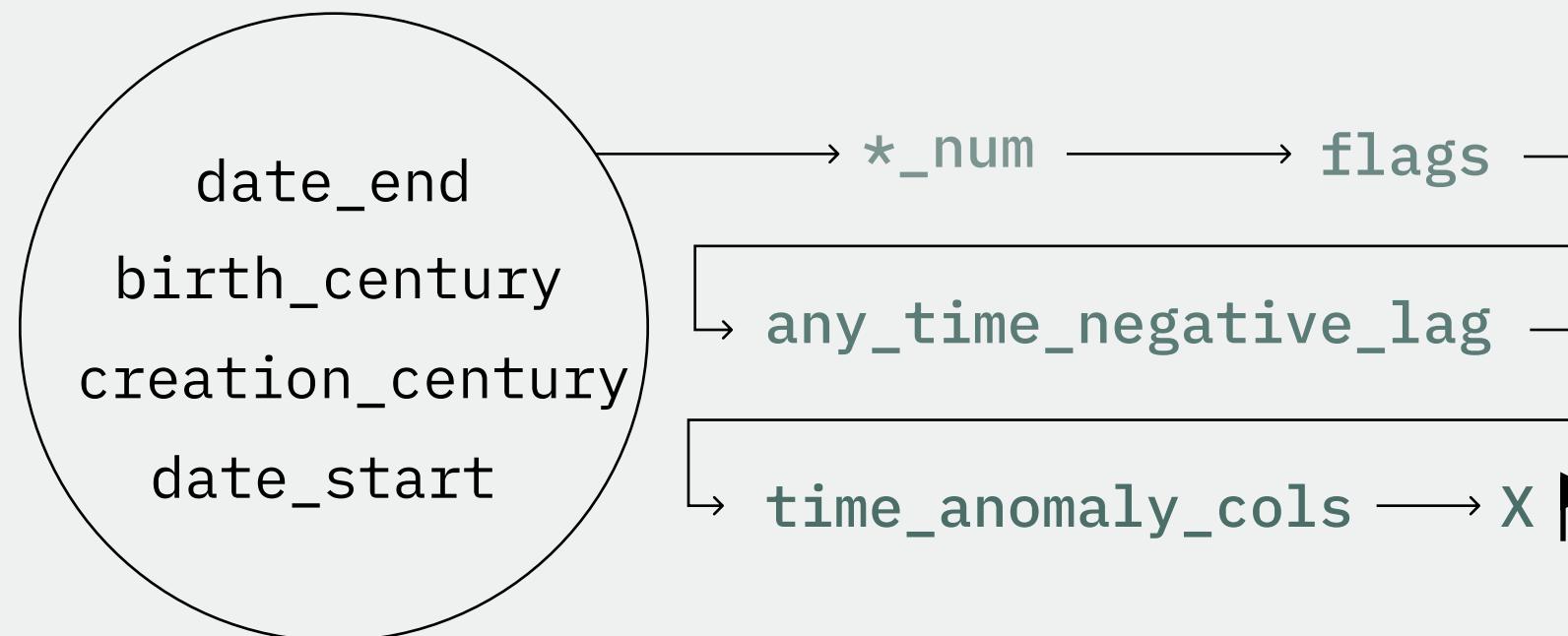
Источник: AIC



ВСЕ  
ЭТИ  
ФИЧИ  
—  
УТЕЧКИ

## Временные и гео- признаки

### KNN clusters для автора



### Дополнительные флаги:

- `flag_early_creation (< 1000)`
- `flag_future_creation (> year_ref)`
- `flag_early_end (< 1000)`
- `flag_future_end (> year_ref)`
- `century_mismatch`

### Дополнительные временные группы:

- `creation decade`
- `half_century_group`
- `century_group`

### KNN признаки

`knn_decade_mean_visible` – сглаженная видимость в окне по десятилетиям. Мягкий сигнал: насколько в для этой эпохи характерна высокая/низкая видимость работ;

`knn_temporal_density` – локальная плотность по дате создания: есть ли вокруг много работ с близкими датами или объект изолирован.

### Новые геопризнаки

`distance_to_museum` – расстояние от работы до музея, рассчитанное по формуле хаверсина, в километрах (центр музея фиксируем как точку в Чикаго);

`knn_geo_density` – географическая плотность



Lighthouse Clock

Simon Willard and Sons

Источник: AIC

# Отбор признаков.

## Модель этапа 2

Surrealist Object  
Functioning Symbolically  
Salvador Dalí  
Источник: АИС



# ОТБОР ФИЧЕЙ

## Финальный feature\_selected

### Удаление утечек

Пирсон,  $\chi^2$ -статистика, ANOVA

### CatBoost

Встроенный отбор

### Lasso

Встроенный отбор

### Фильтры

Пирсон,  $\chi^2$ -статистика, ANOVA

### RFECV

#	Feature	Importance
1	has_multimedia_resources_extra	8.493
2	department_te	7.156
3	thumbnail_area	5.933
4	knn_mean_visible	5.568
5	age_at_creation	4.932
6	classification_title	4.603
7	knn_rarity_density	4.493
8	side_normalised	3.534
9	knn_author_cluster	2.707

## BASELINE. CatBoostClassifier

depth=6, lr=0.1, loss\_function="Logloss"

**0.93** ROC AUC

ст. отклонение на KFold = 0.0032

**0.60** F1

ст. отклонение на KFold = 0.0042

**0.93** PRECISION

**0.60** RECALL

(134078, 84)

## МОДЕЛЬ ЭТАПА 2. CATBOOST + 5-FOLD CV

параметры те же

**0.925** ROC AUC

ст. отклонение на KFold = 0.0004

**0.58** F1

ст. отклонение на KFold = 0.0042

**0.55** PRECISION

**0.67** RECALL

(134078, 102)



# Интерпретация финальной модели

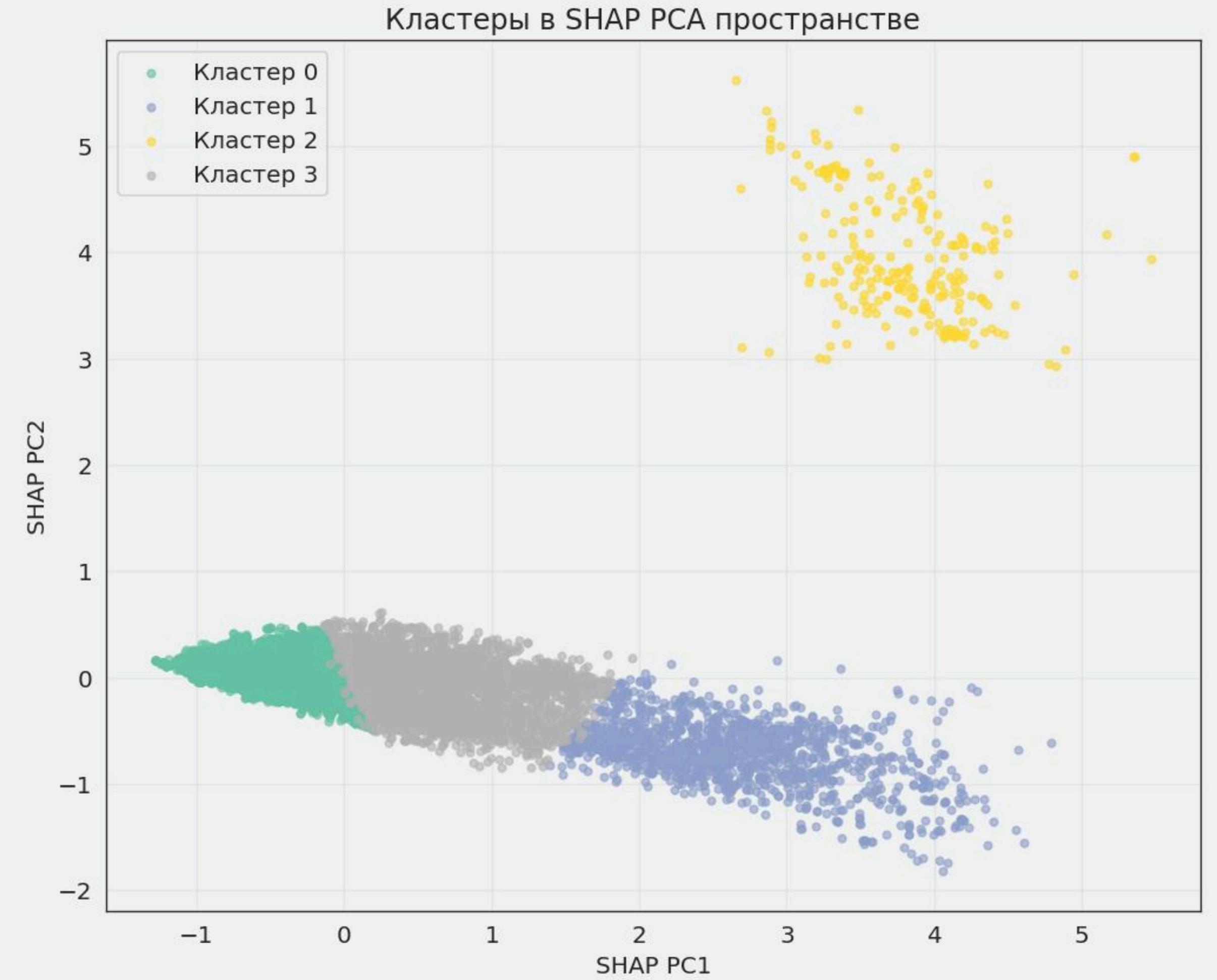
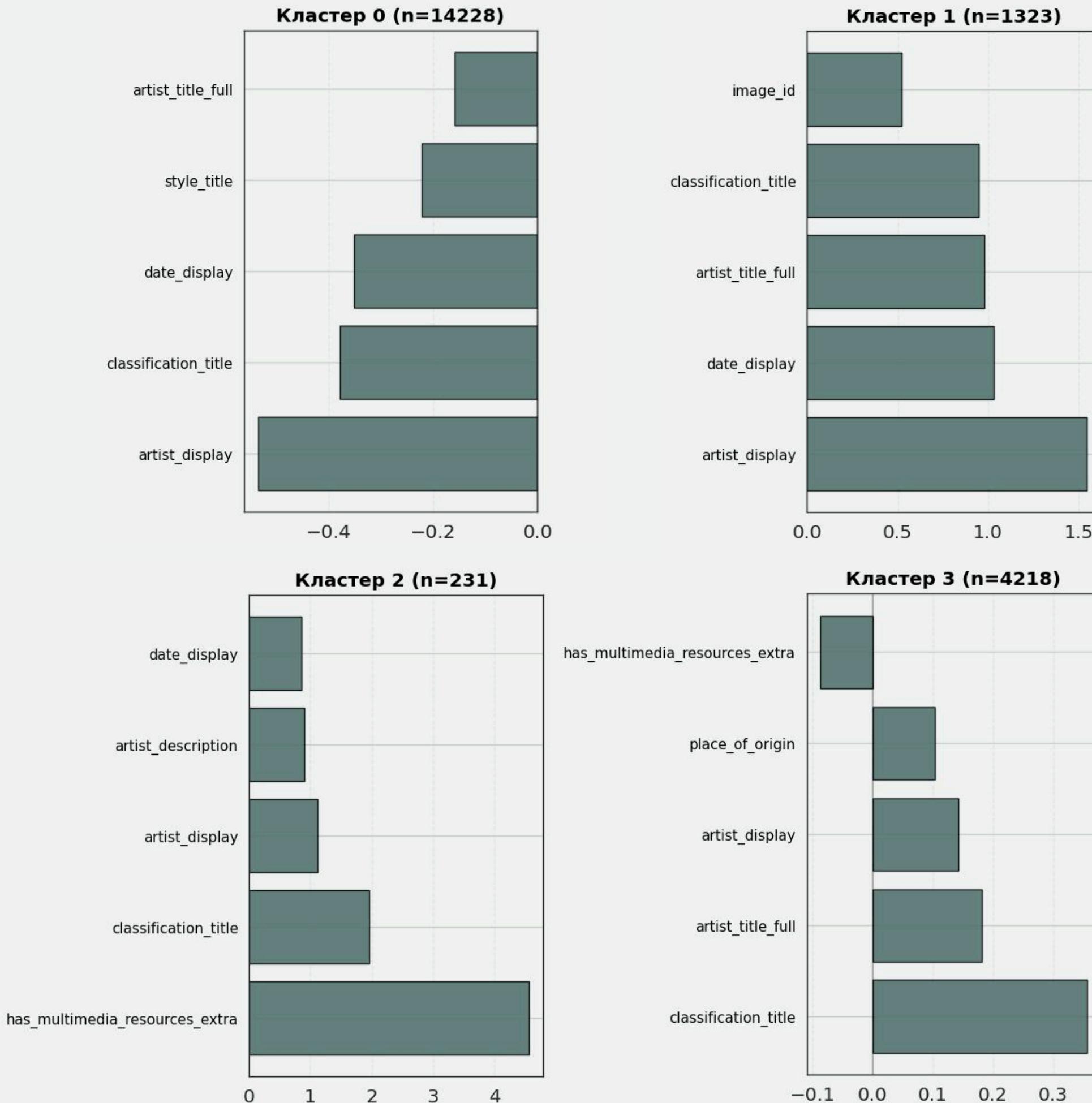


Buddha Shakyamuni Seated  
in Meditation

Tamil Nadu

Источник: АИС

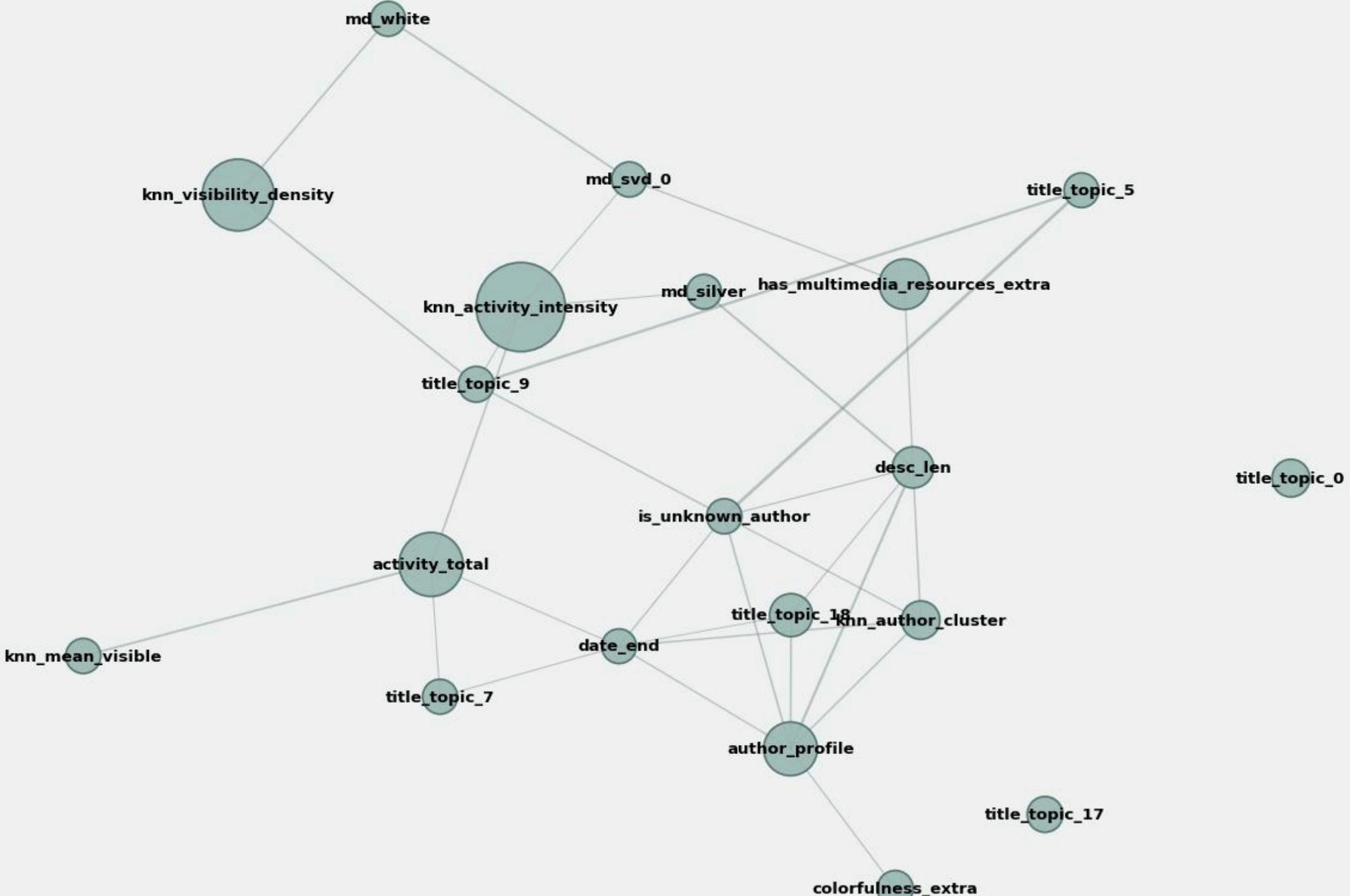
# SHAP-ЭМБЕДДИНГИ НА 20000. КЛАСТЕРЫ РЕЖИМОВ ВИДИМОСТИ



# SHAPELY FLOW. КАКИЕ СВЯЗКИ РАБОТАЮТ?

МЫСЛИМ КАК КУРАТОР

## Shapley Flow без лиц: ядро совместных влияний признаков



1

## Авторско-исторический блок

Когда у работы мало “современных” сигналов (слабая цифровая оболочка, мало описаний), модель вынуждена опираться на то, что можно назвать историческим статусом.

2

## Художественно-объектный блок

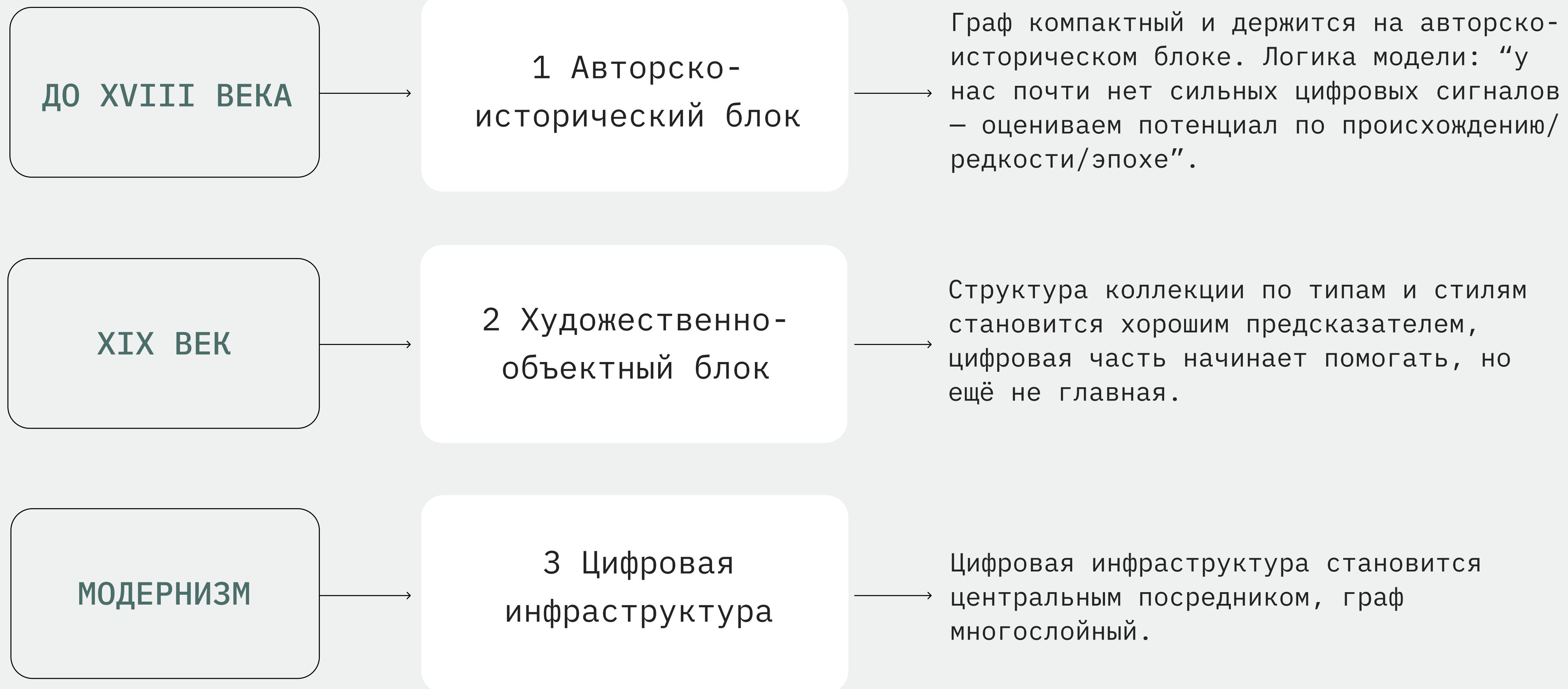
Что это как объект искусства?

3

## Цифровая инфраструктура

Насколько объект подготовлен к жизни в цифровых каналах

# SHAPELY FLOW. КАК МЕНЯЕТСЯ РЕЖИМ ПО ЭПОХАМ

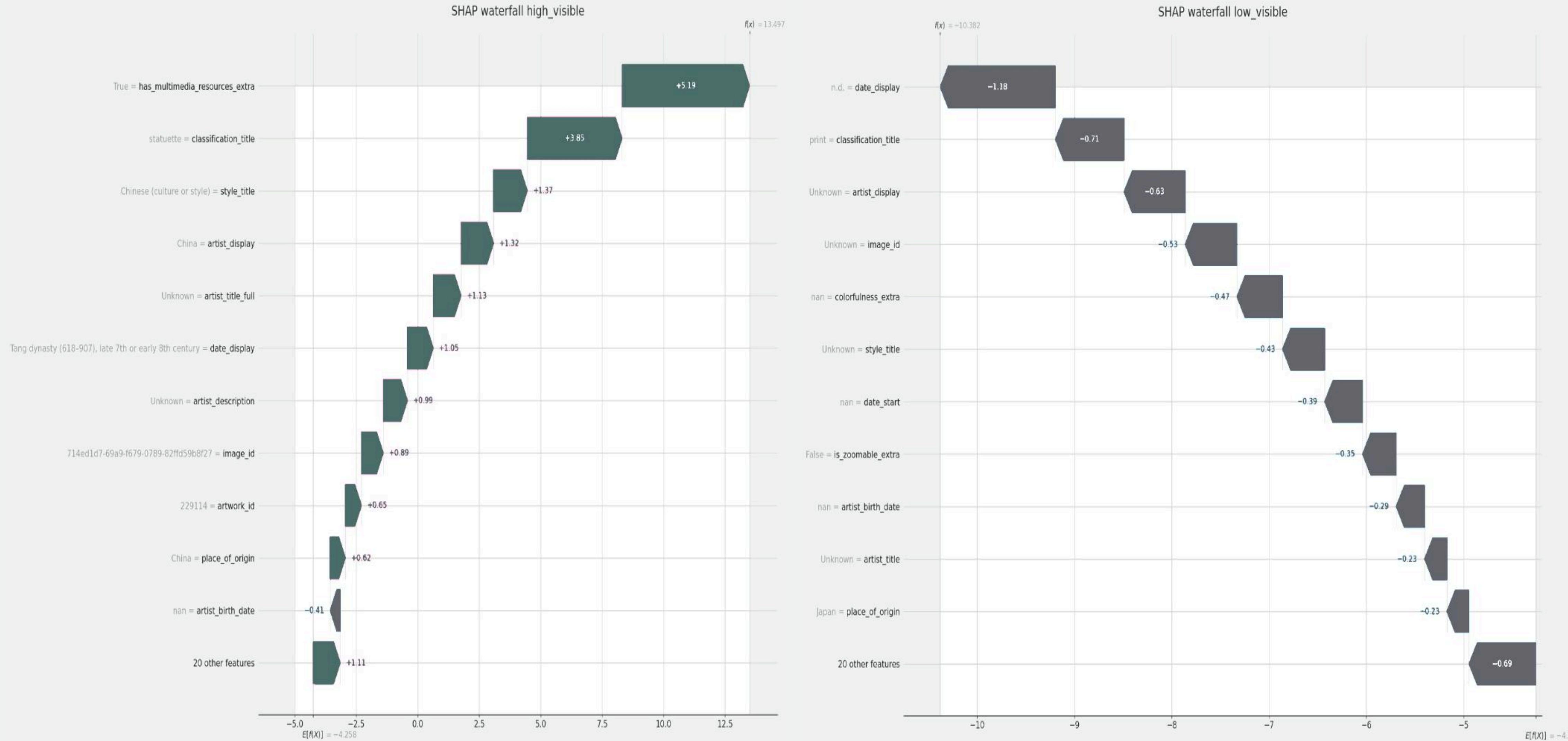


## ВЫВОД ПО SHAPLEY FLOW

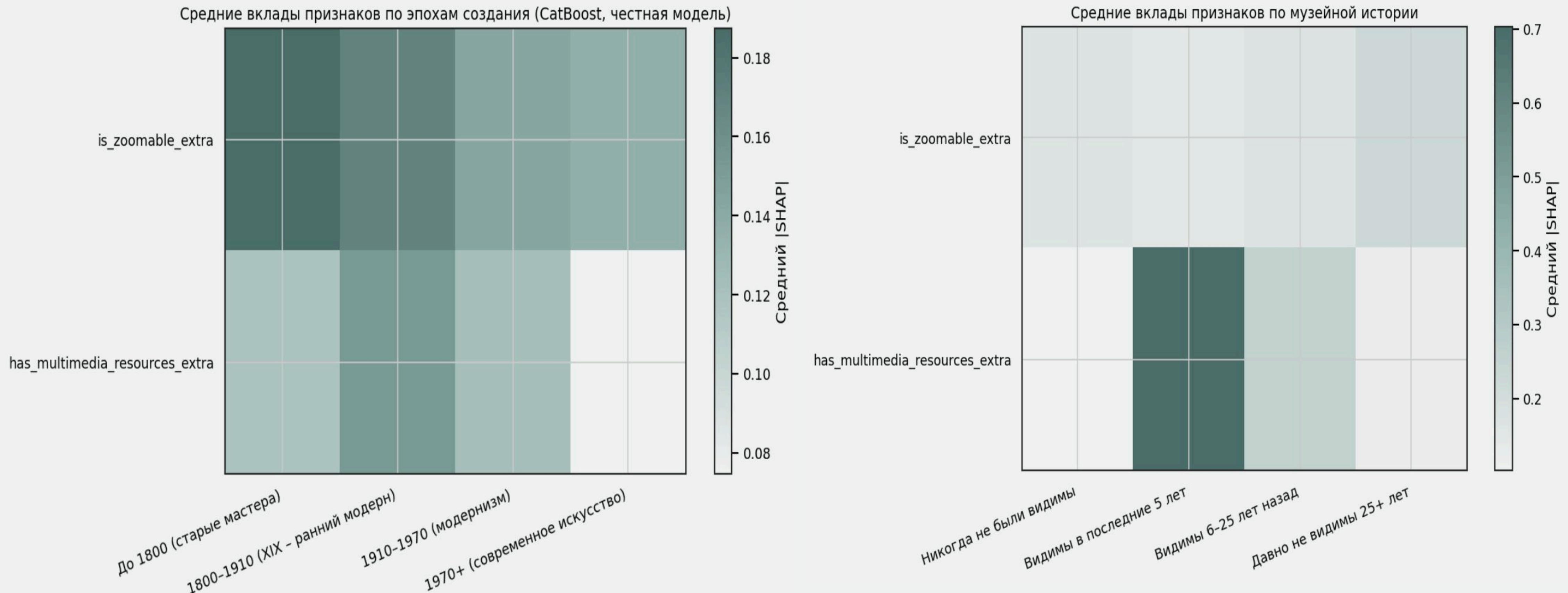
Shapley Flow показывает конкретные связи признаков, которые выводят работу в видимость – и подсказывает, что именно усилить (цифру, контекст или подачу), чтобы она попала в экспозицию и продукты.

The Fountain, Villa Torlonia,  
Frascati, Italy, 1930  
John Singer Sargent  
Источник: AIC





# SHAP ПО СЕГМЕНТАМ КОЛЛЕКЦИИ



# СРАВНЕНИЕ ТРЕХ МОДЕЛЕЙ

Model A X_base	Model B Эмбеддинги + SHAP-кластеры	Model C X_base + SHAP-эмбеддинги + cluster	Model D Логистическая регрессия по X_base
<b>0.90</b> AUC <b>0.51</b> F1  Это качество, которое мы получим в продакшне <b>при прогнозе "на будущее"</b> .  F1 указывает на сложность балансировки precision/recall - модель осторожничает	<b>1.0</b> AUC <b>1.0</b> F1  SHAP-значения (и кластеры) содержат полную информацию о целевой переменной.  Модель выучила детерминированные правила, разделяющие объекты.  <b>Использовать для прогноза нельзя (data leakage)</b> , но идеальны для <b>**аналитики**</b> . Если мы знаем, в какой SHAP-кластер попадает объект, мы 100% знаем его судьбу		<b>0.68</b> AUC <b>0.25</b> F1  Плохое качество модели.  Мы решили не использовать ее в финальном анализе и предсказаниях.

# ГЛАВНЫЕ НАХОДКИ

## "СПЯЩИЙ ГИГАНТ"

Около **92%** объектов коллекции попадают в кластер 0 с низкой структурной совместимостью с историческими паттернами видимости. Модель устойчиво относит эти объекты к области данных, где в прошлом отсутствовала активность

## РЕЦЕПТ ВИДИМОСТИ

Для попадания в активные 8% (Кластеры 1-5) недостаточно просто "быть шедевром". Необходим **"цифровой след"** (мультимедиа) или **"научный след"** (публикации).

## КАЧЕСТВО ДАННЫХ

Модель нашла **четкие паттерны**. Ошибки модели часто соответствуют объектам, находящимся на границе распределений, что указывает либо на шум и неполноту данных, либо на структурные несоответствия между признаками и историческими лейблами.



**Study for "Victory"**  
Evelyn Beatrice Longman  
Источник: AIC

КОНЕЦ.  
Ждем ваши вопросы

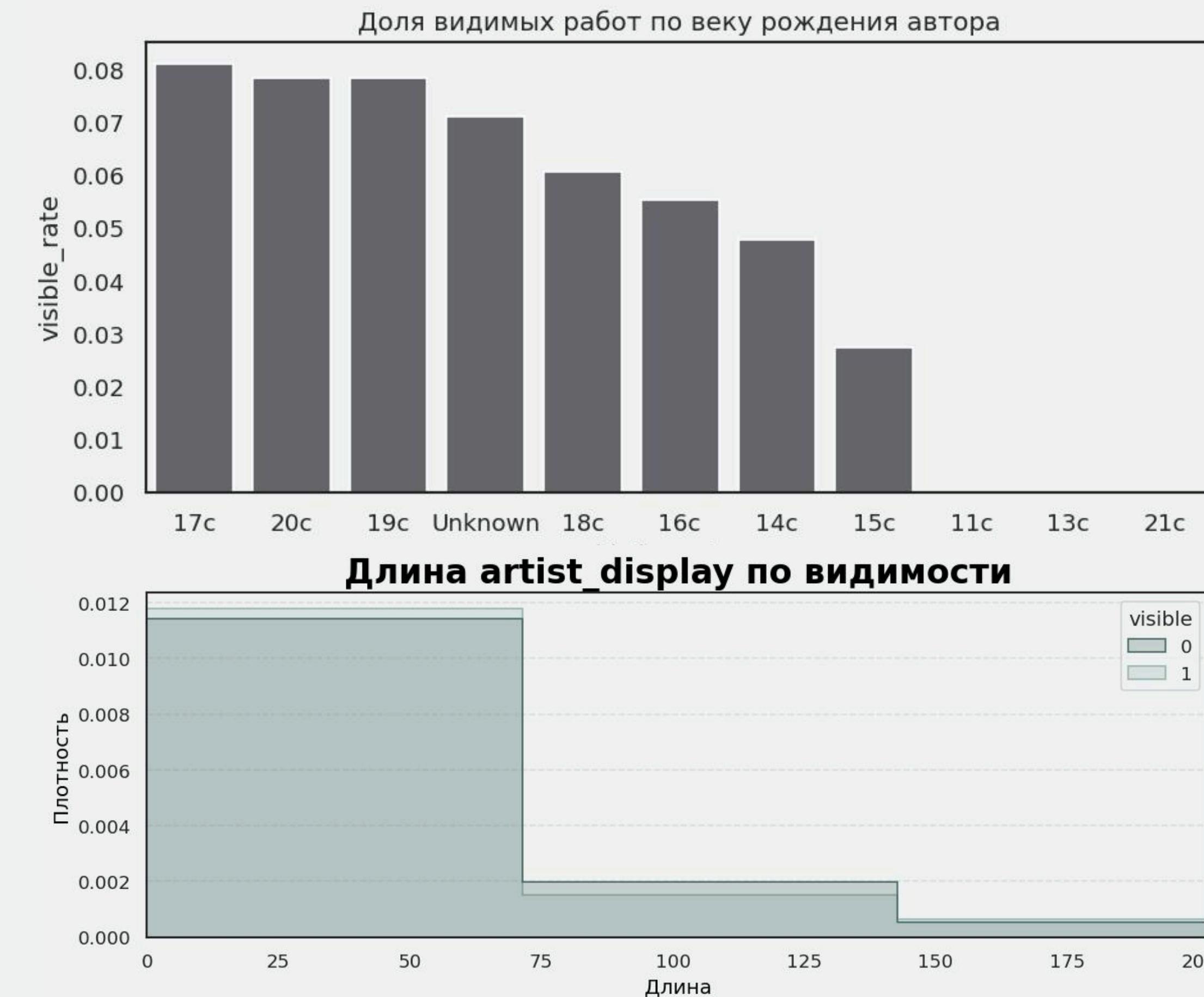
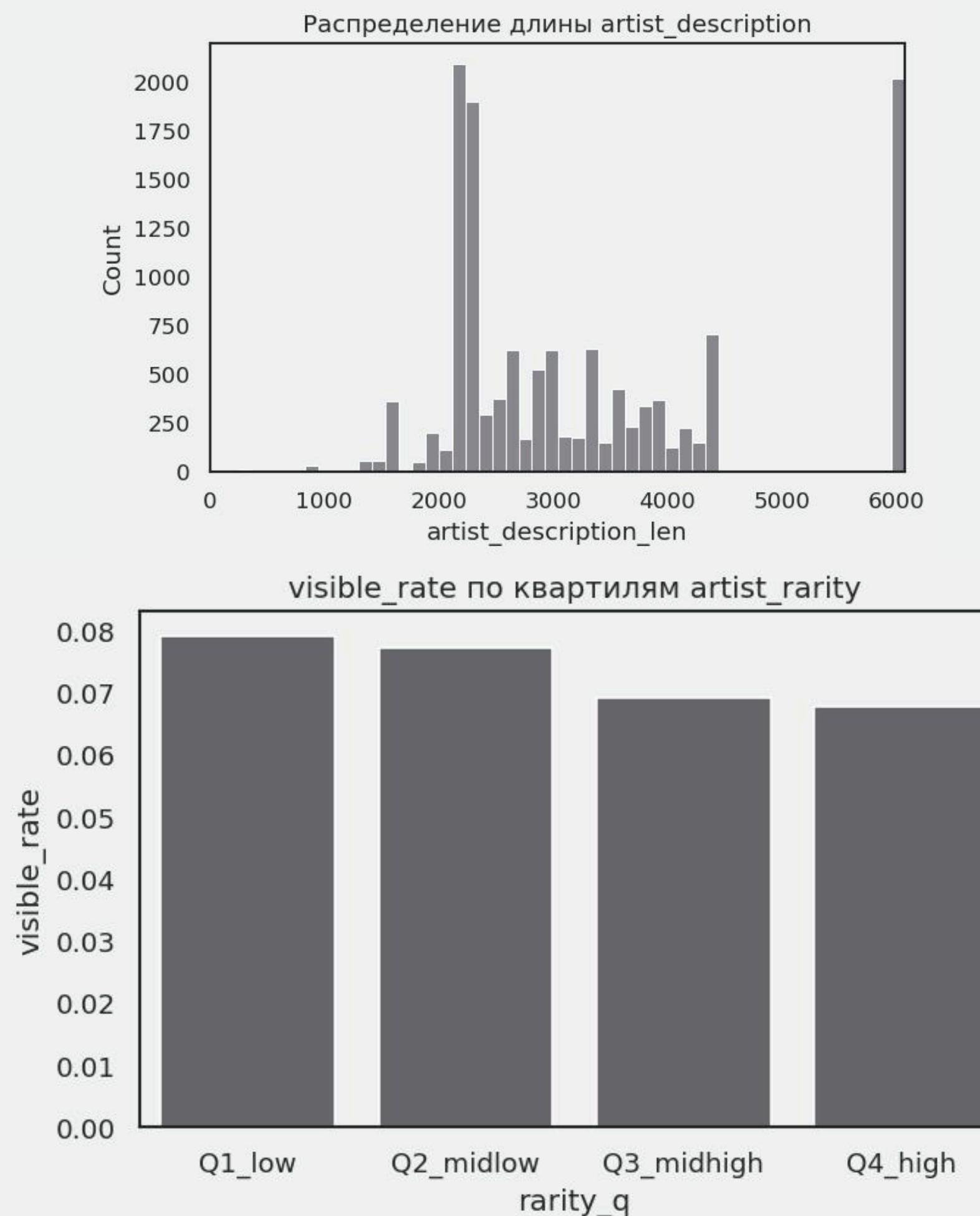
Sally Slips Bye-Bye, 1972  
Jim Nutt  
Источник: АИС



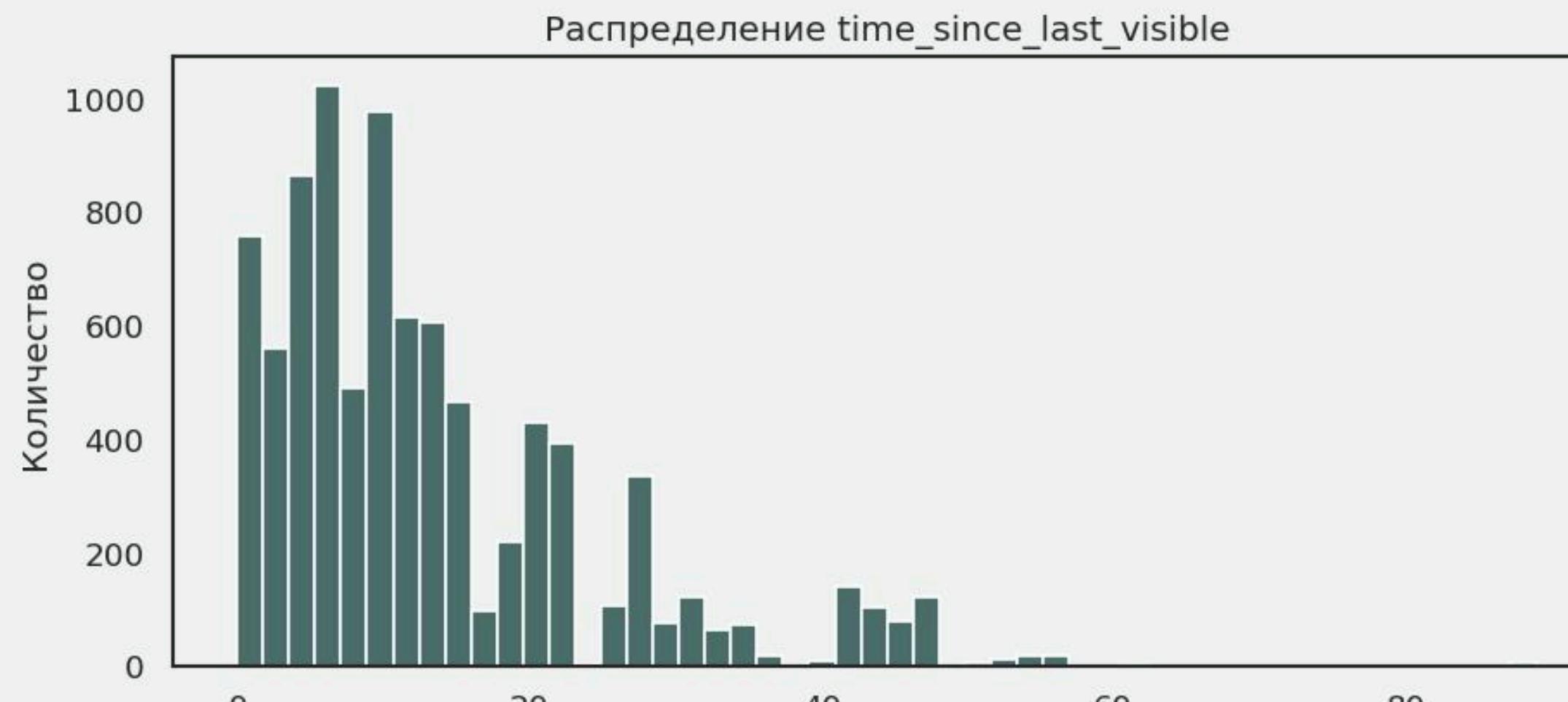
# Приложение

---

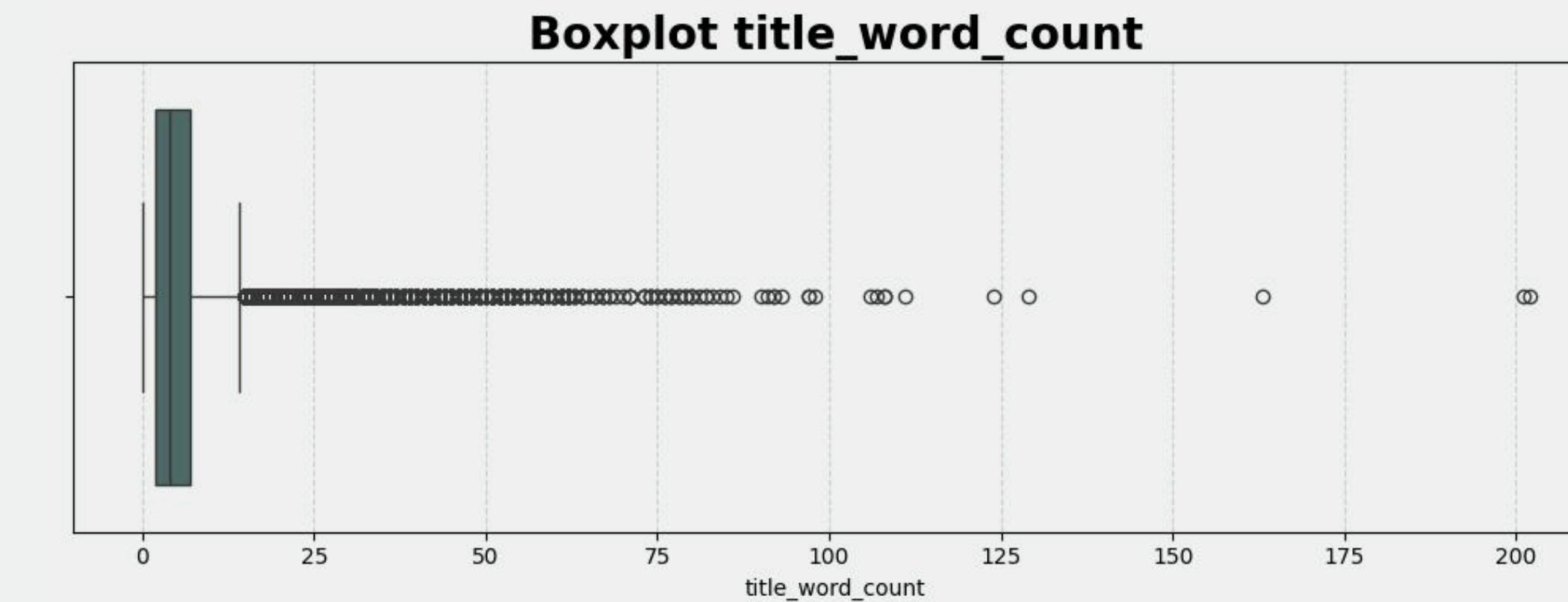
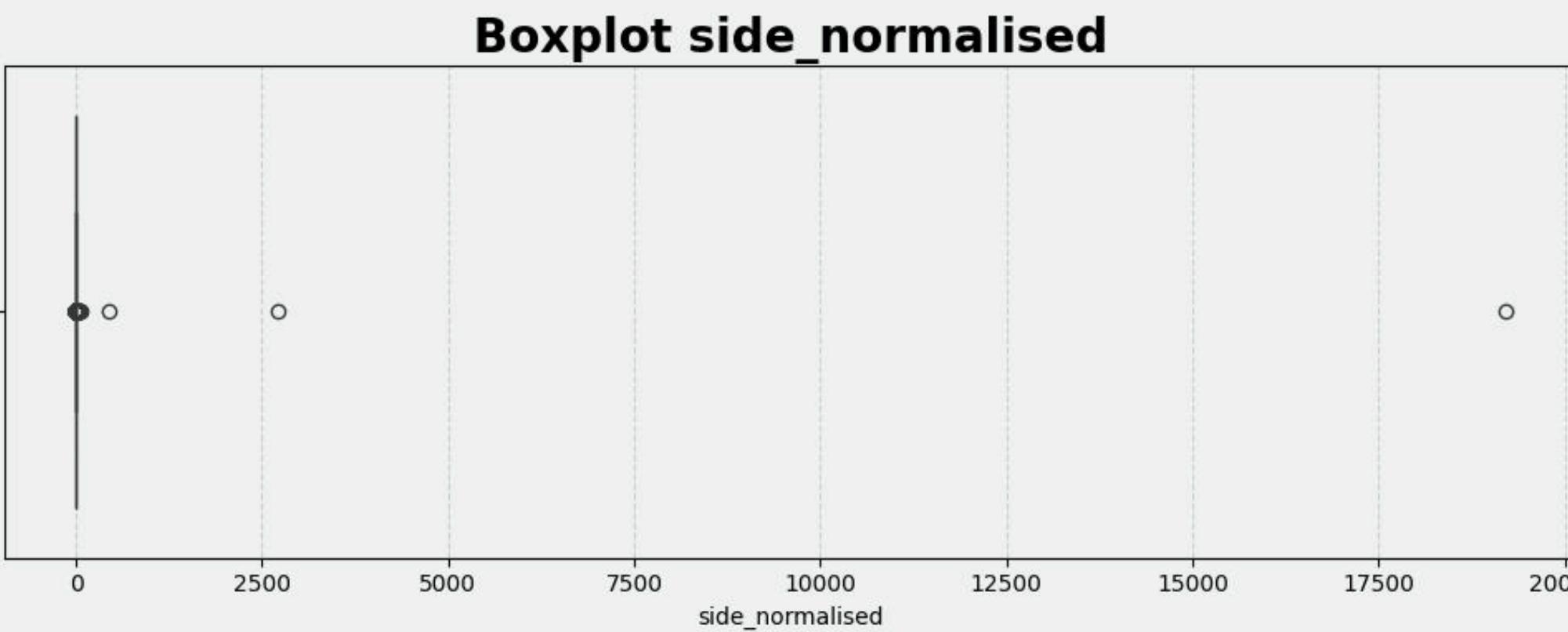
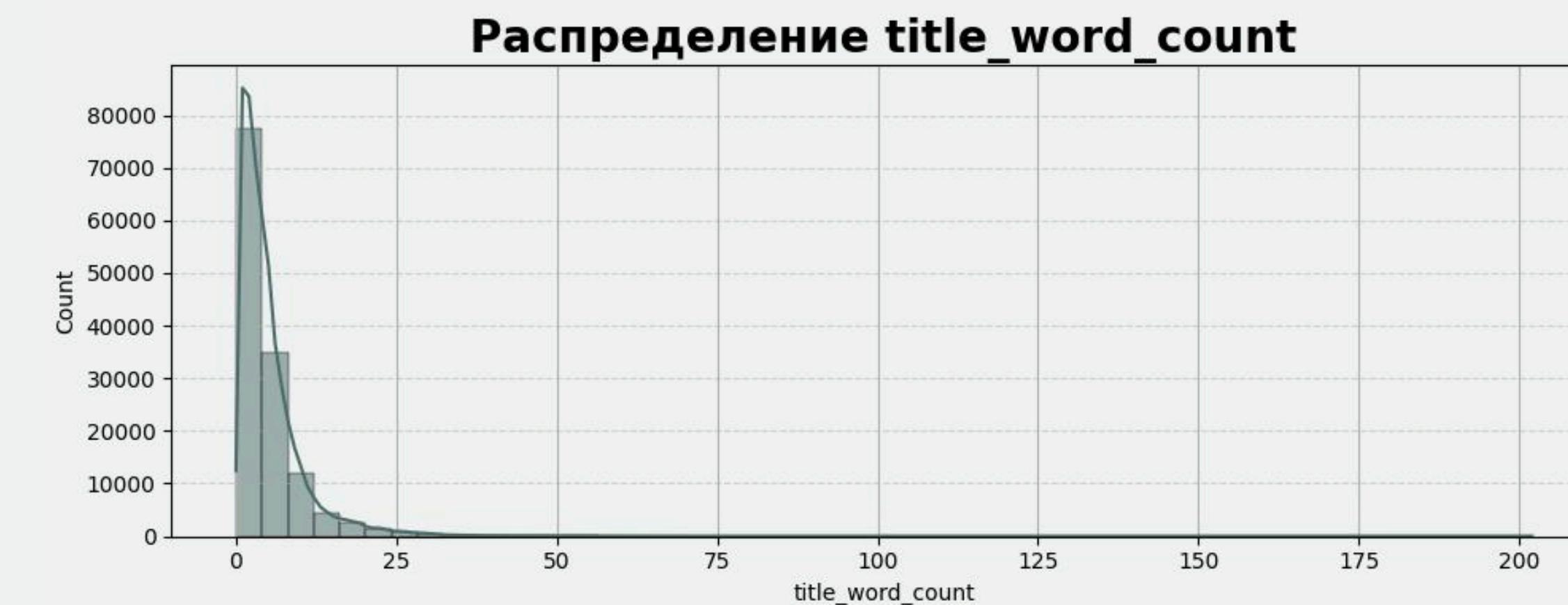
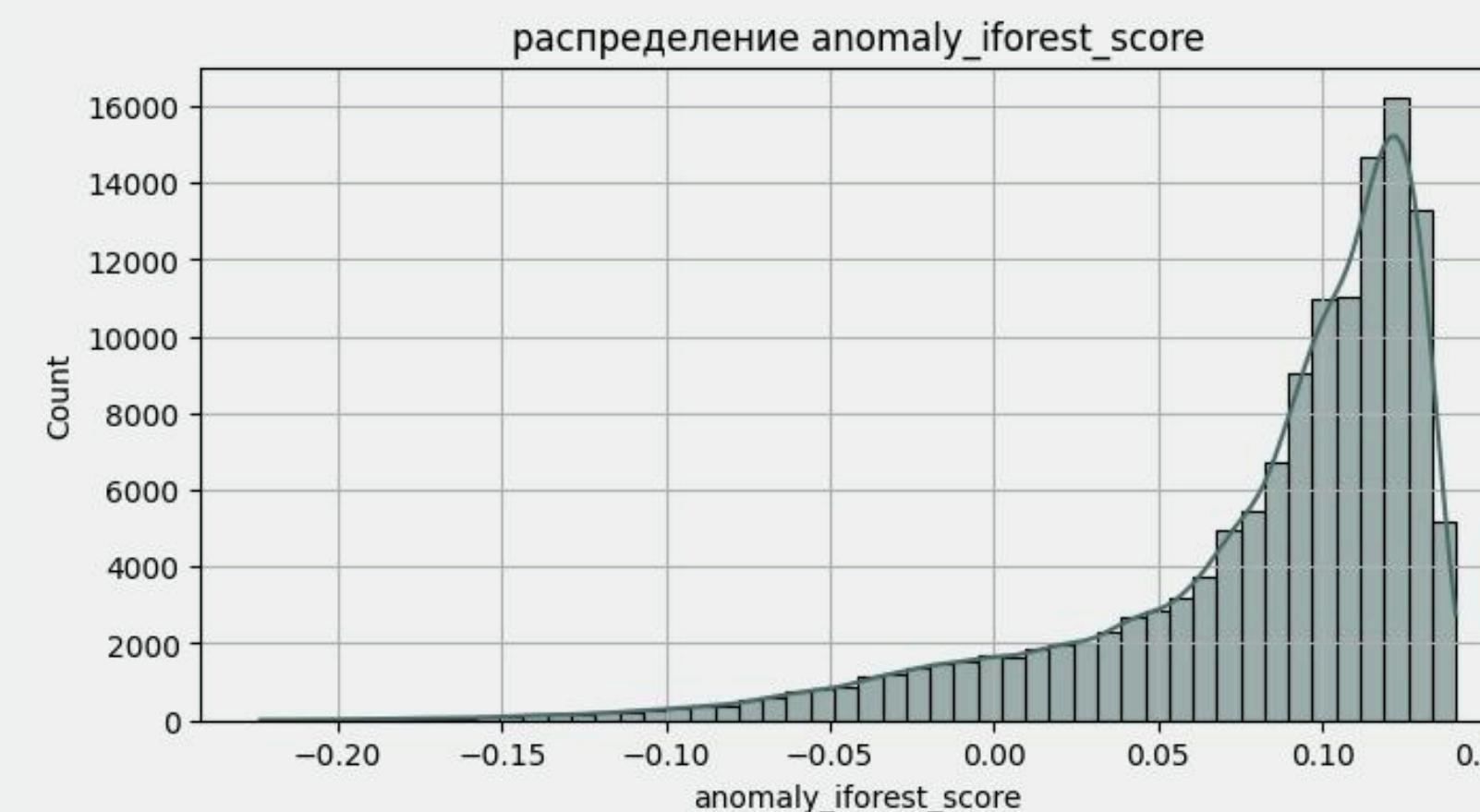
# ДОПОЛНИТЕЛЬНЫЕ ГРАФИКИ ПО АВТОРСКИМ ПРИЗНАКАМ (ДЛИНА ОПИСАНИЯ, ВЕК РОЖДЕНИЯ, РЕДКОСТЬ)



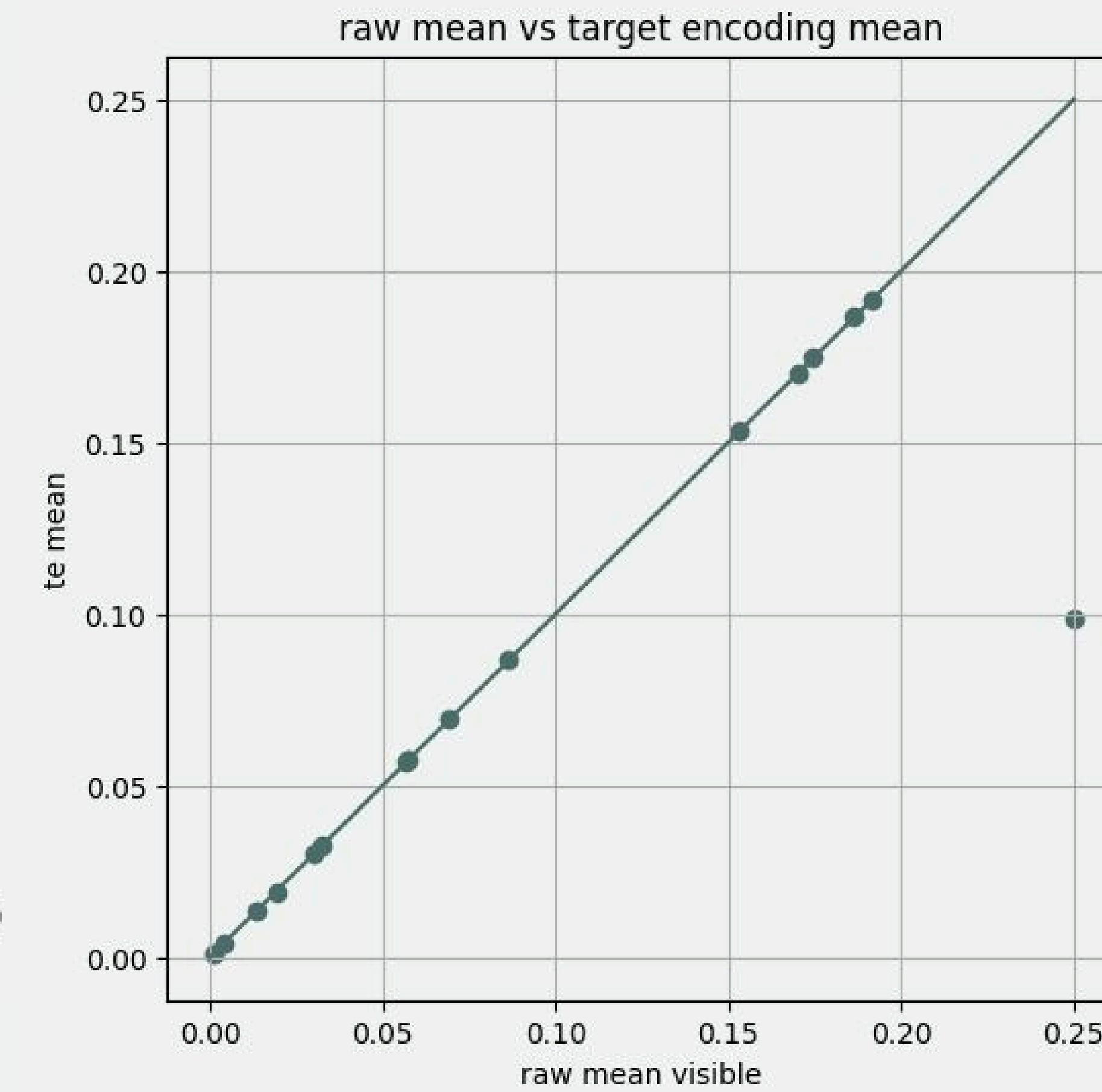
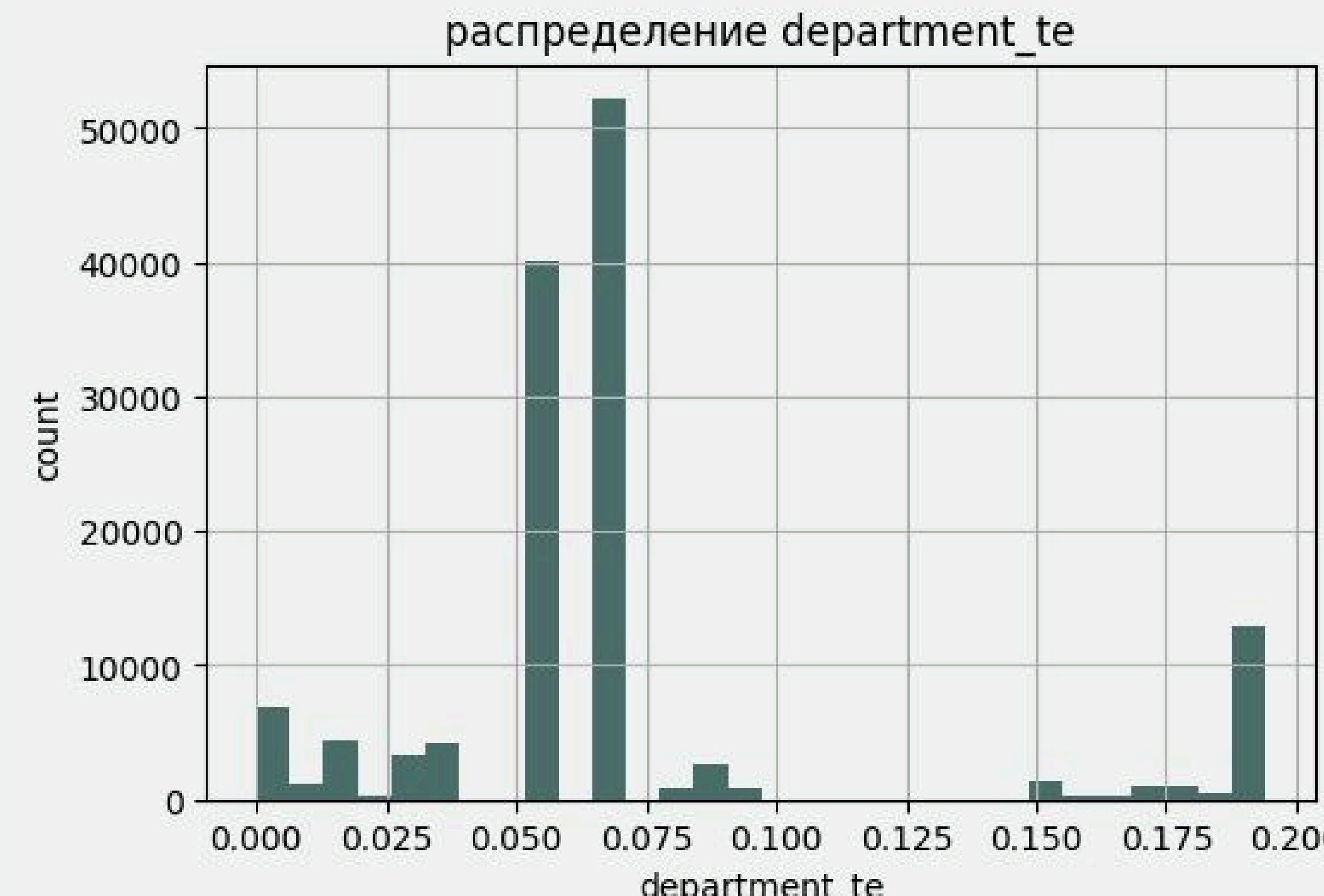
# ДОПОЛНИТЕЛЬНЫЕ ГРАФИКИ ПО ВРЕМЕННЫМ ПРИЗНАКАМ (ВРЕМЯ С ПОСЛЕДНЕЙ ВЫСТАВКИ, ВОЗРАСТ РАБОТЫ)



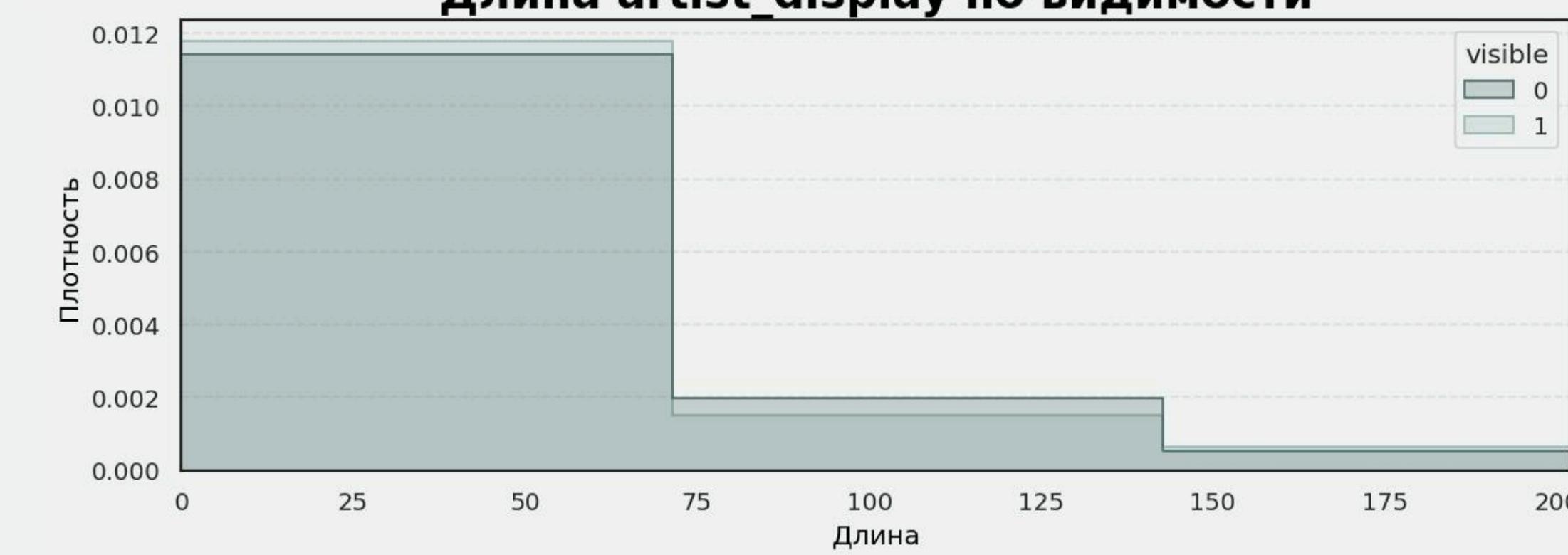
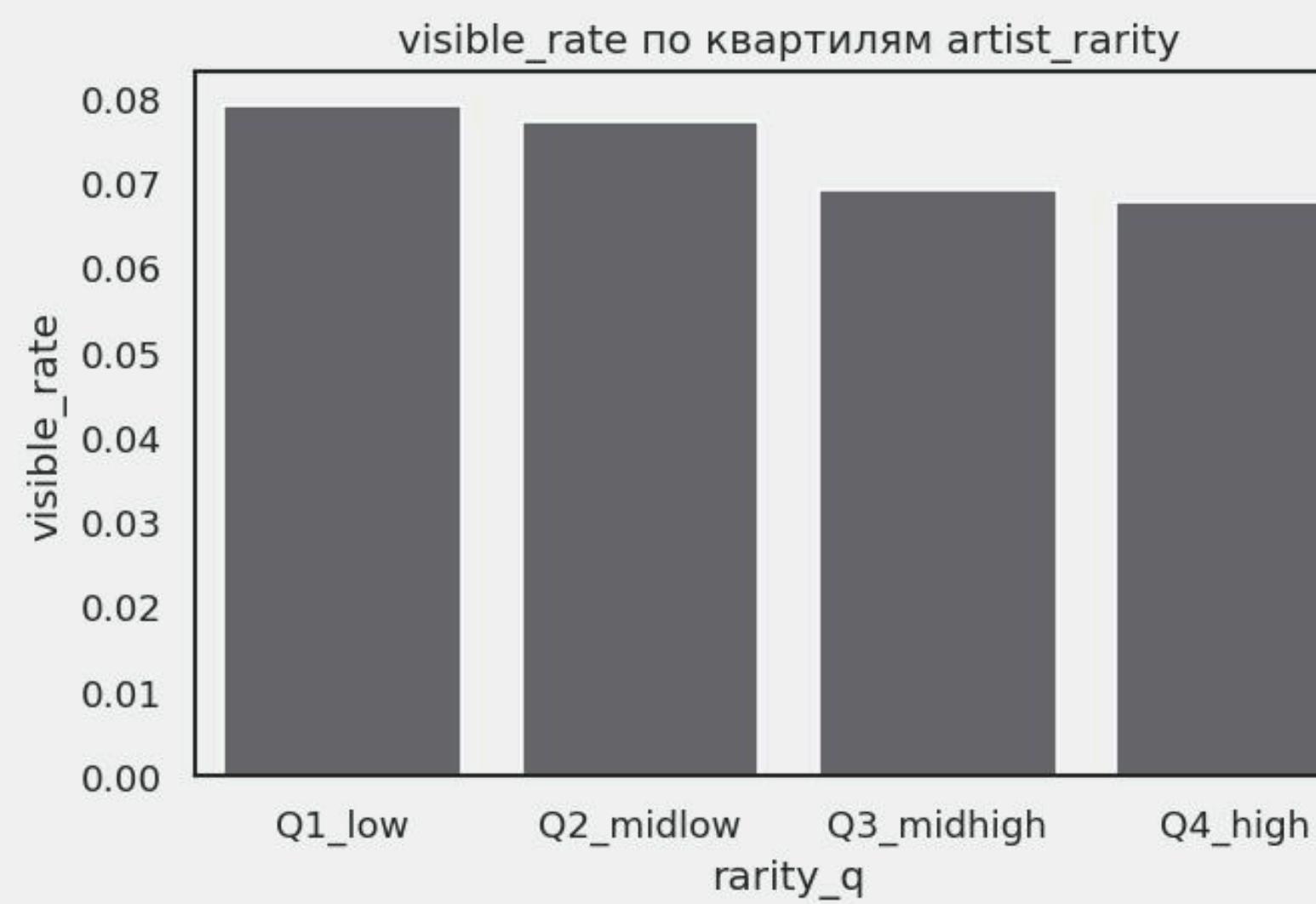
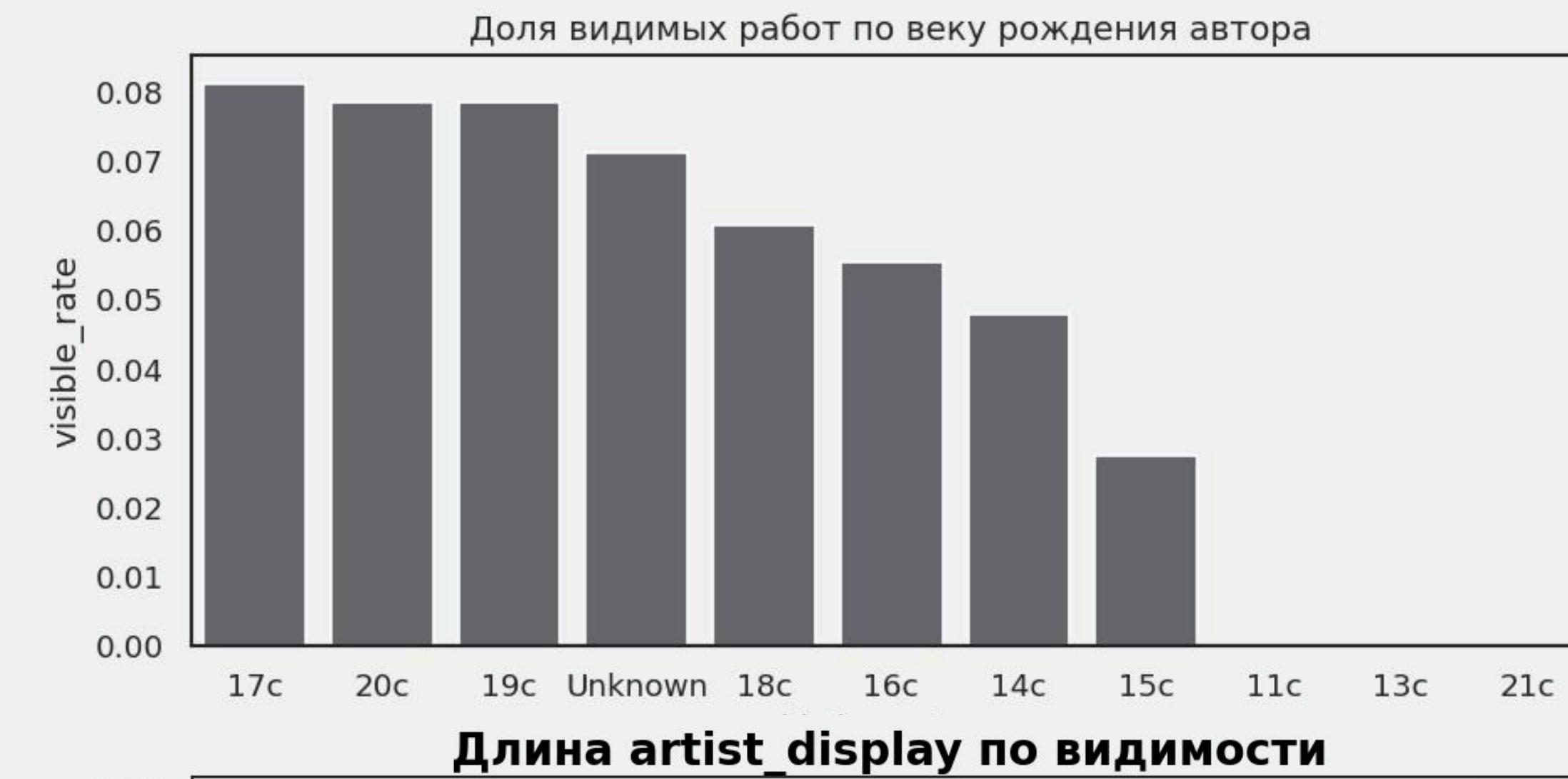
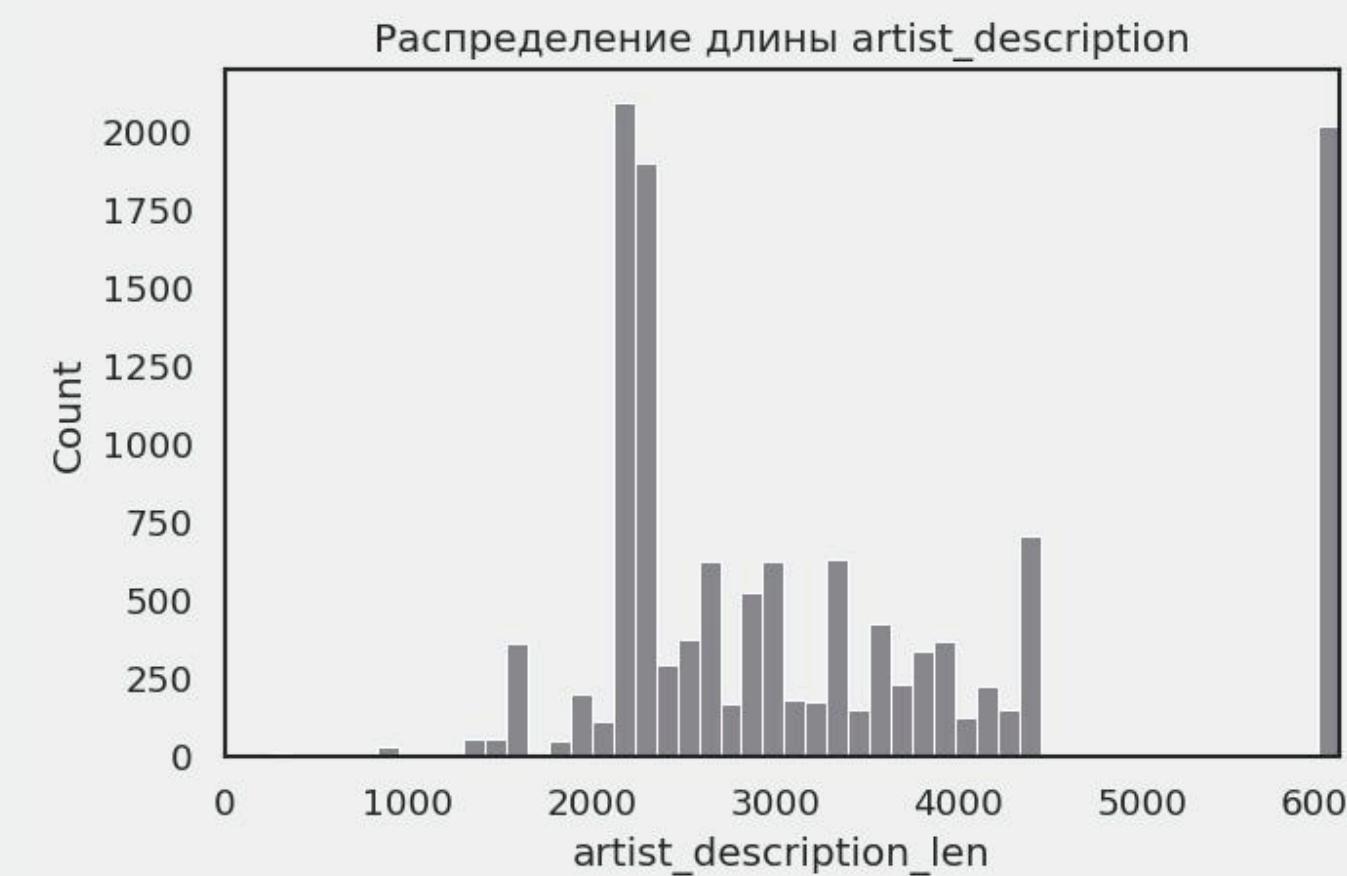
# ДОПОЛНИТЕЛЬНЫЕ ГРАФИКИ ПО ХУД. ПРИЗНАКАМ (IFO, TITLE\_WORD\_COUNT, SIDE\_NORMALISED)



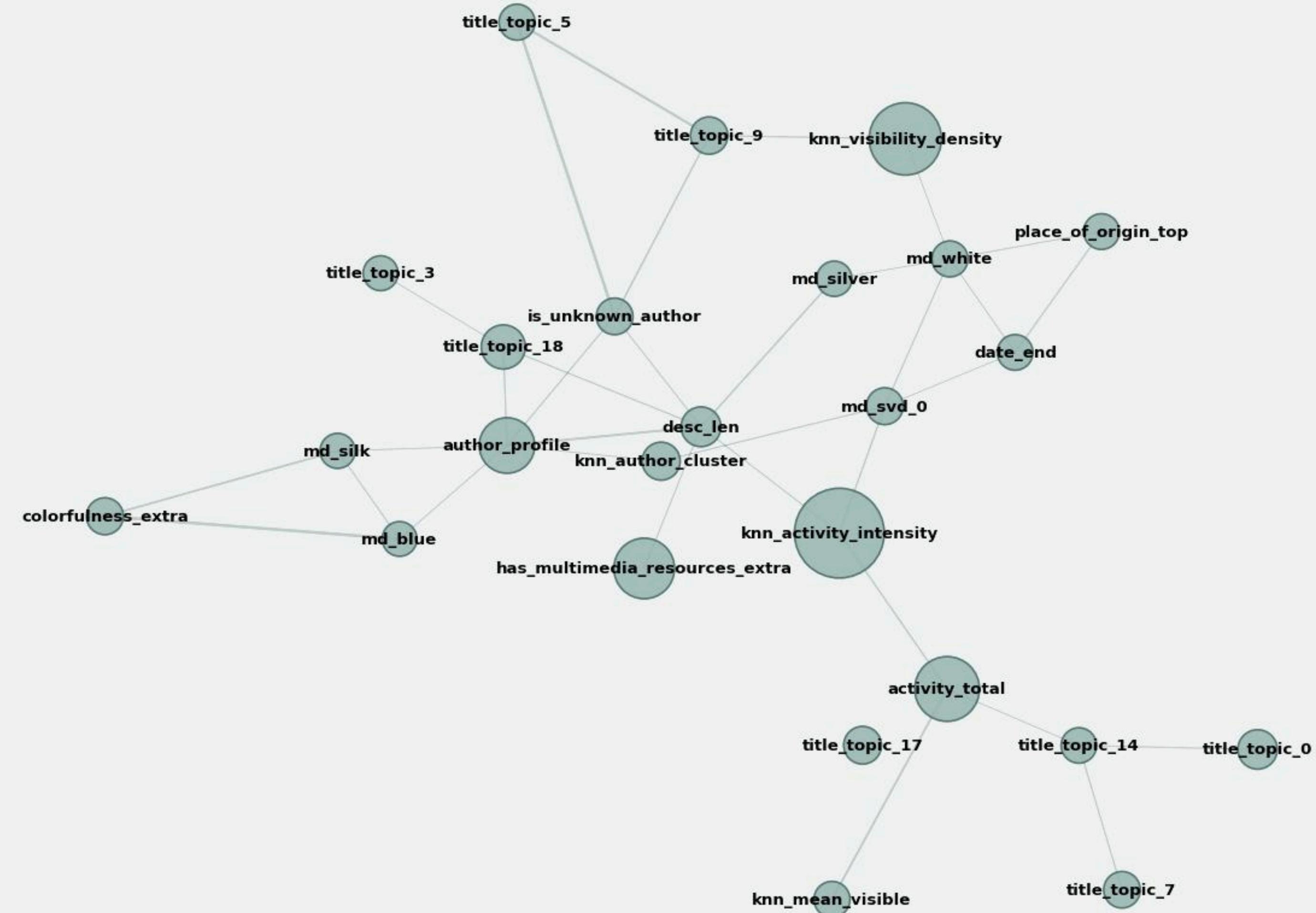
# ДОПОЛНИТЕЛЬНЫЕ ГРАФИКИ ПО ХУД. ПРИЗНАКАМ (DEPARTMENT\_TE)



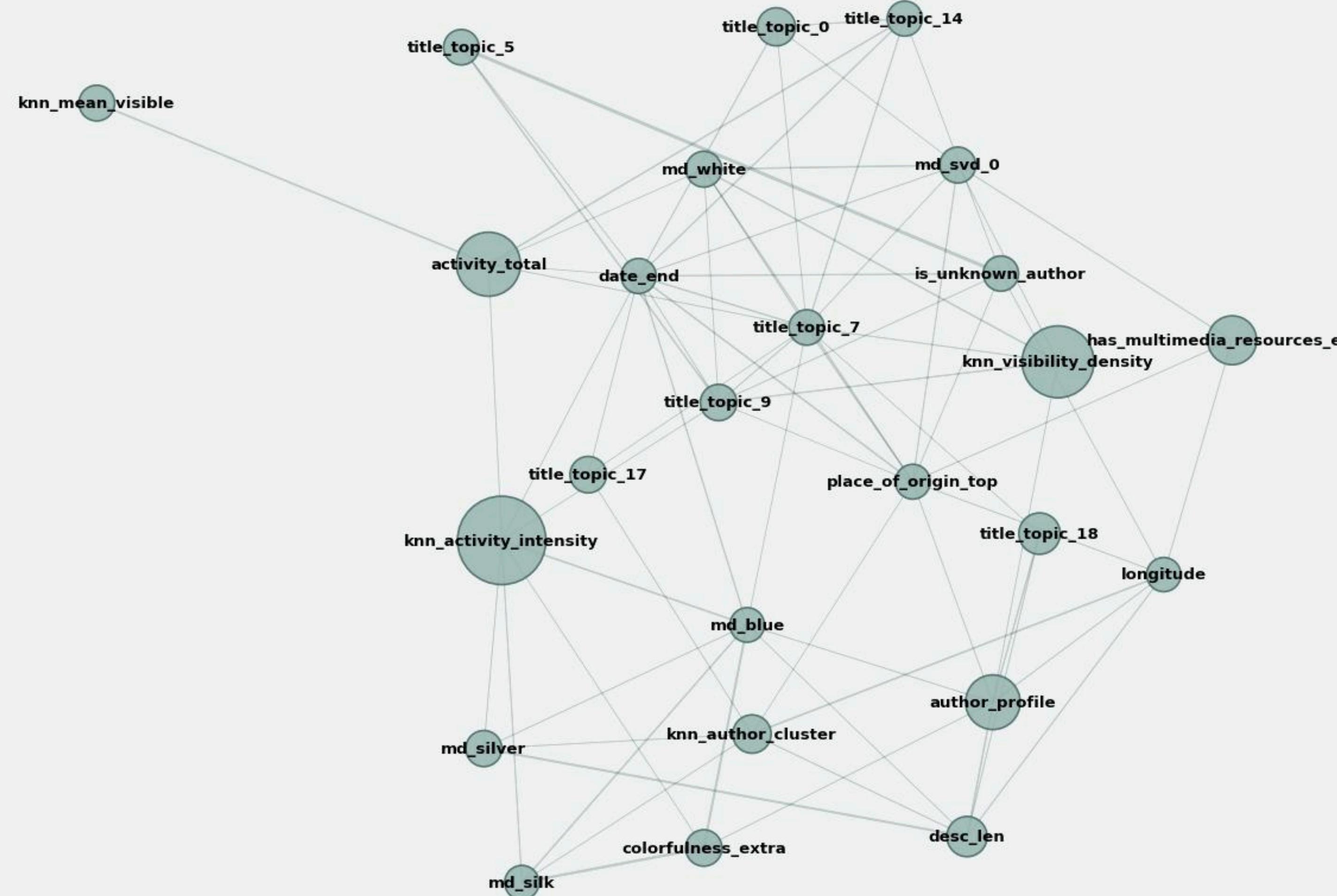
# ДОПОЛНИТЕЛЬНЫЕ ГРАФИКИ ПО АВТОРСКИМ ПРИЗНАКАМ (ДЛИНА ОПИСАНИЯ, ВЕК РОЖДЕНИЯ, РЕДКОСТЬ)



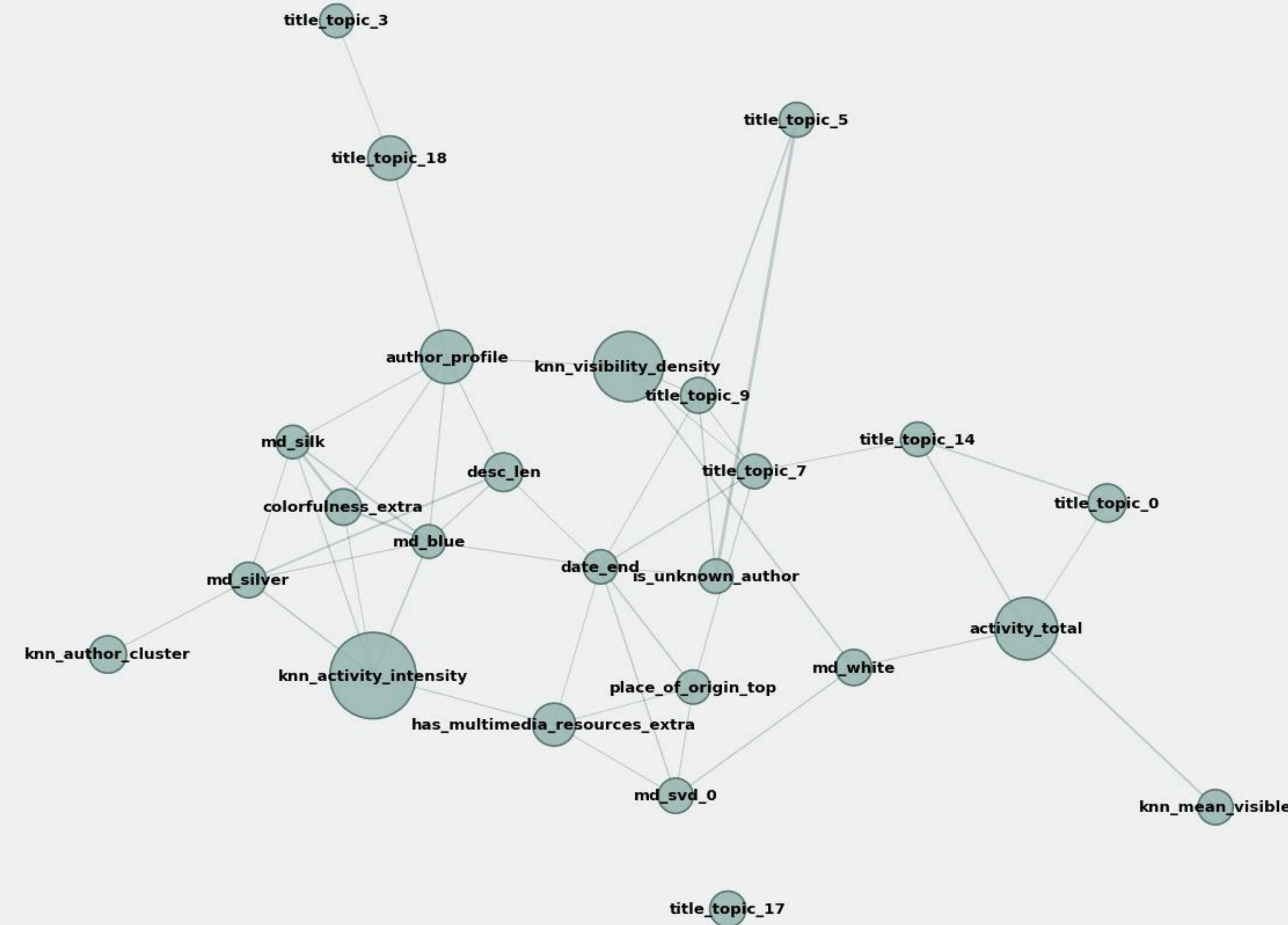
## Shapley Flow: Создание 1800-1910 (XIX - ранний модерн) (n=5156)



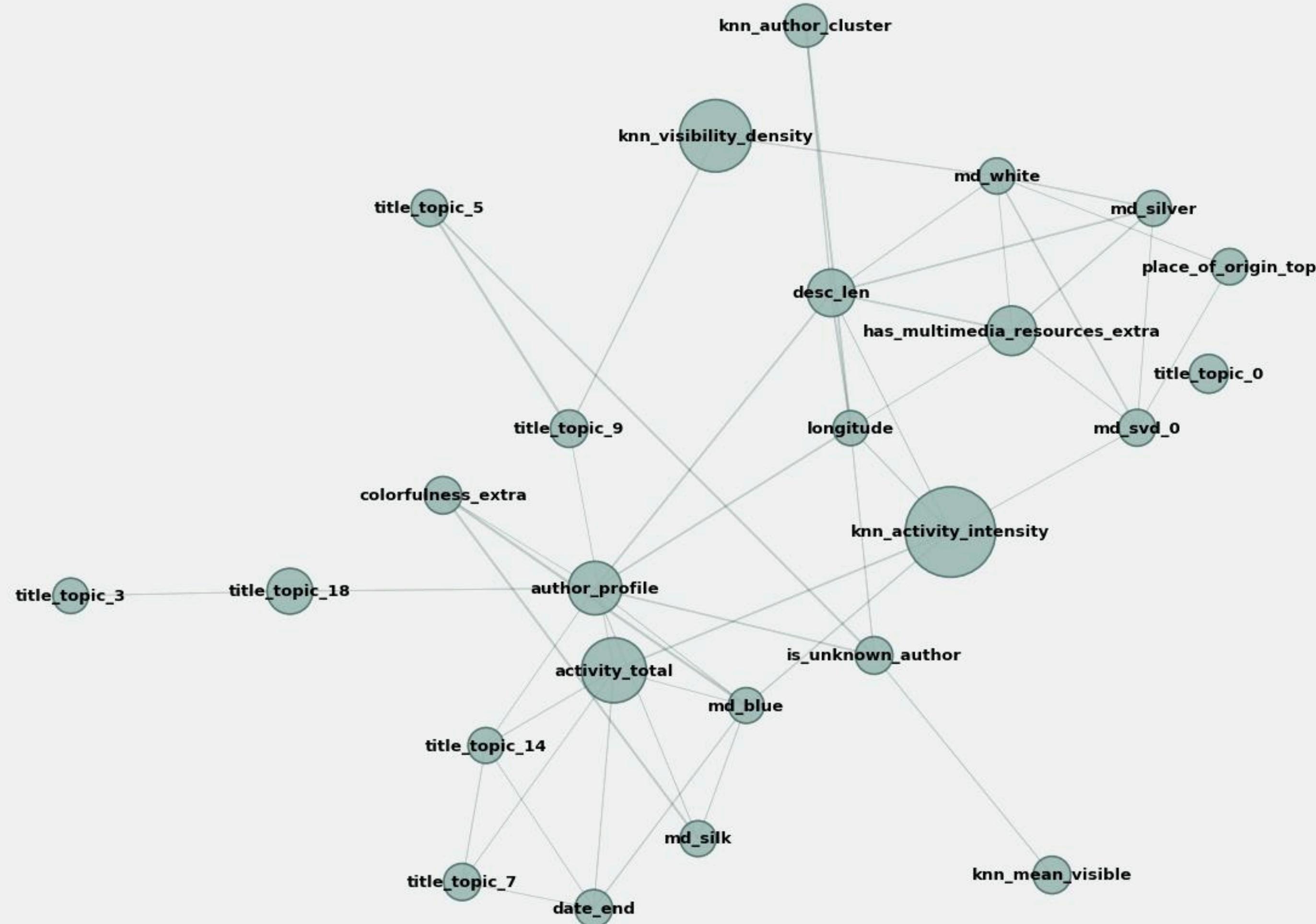
## Shapley Flow: Создание 1910-1970 (модернизм) (n=5527)



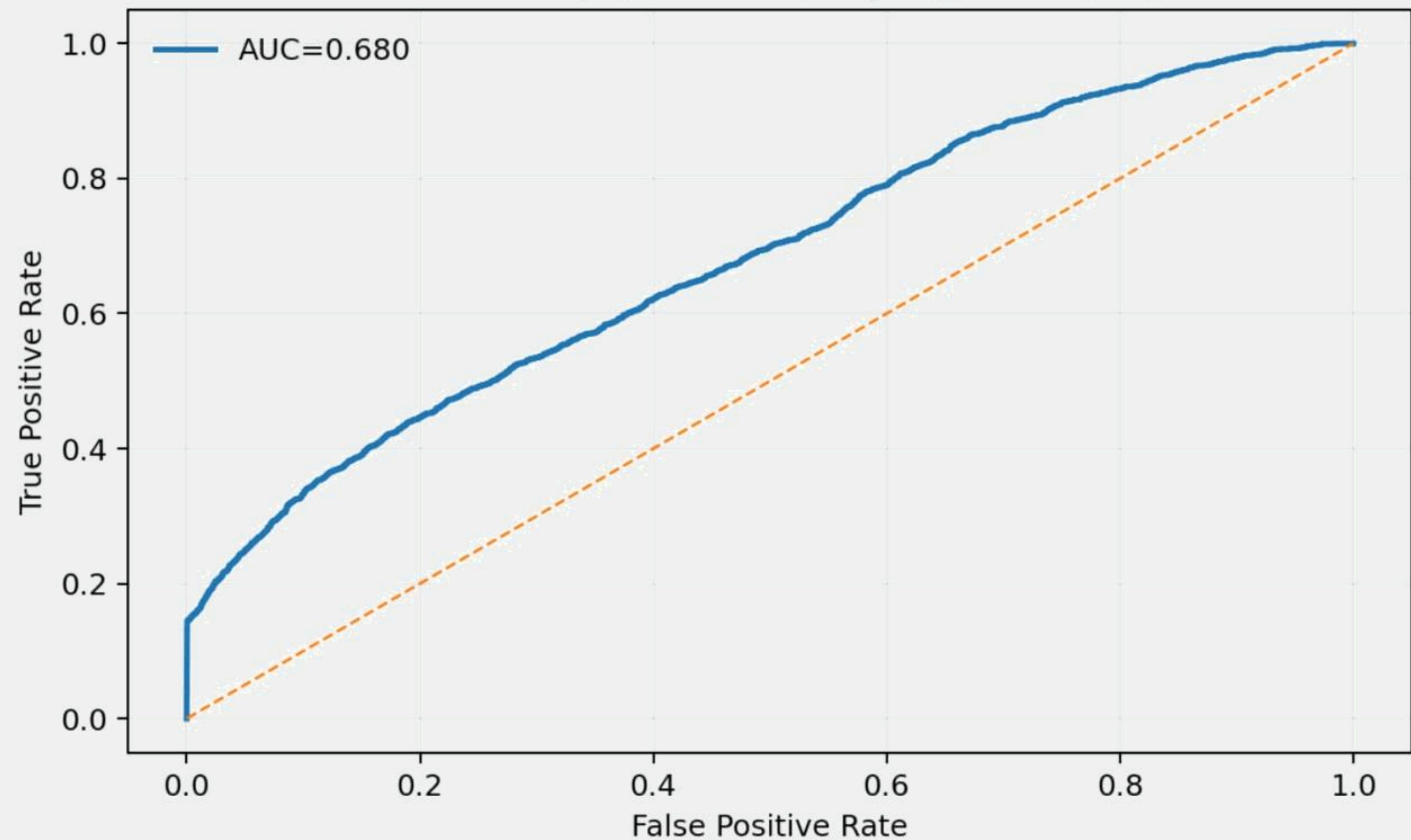
## Shapley Flow: Создание 1970+ (современное искусство) (n=3140)



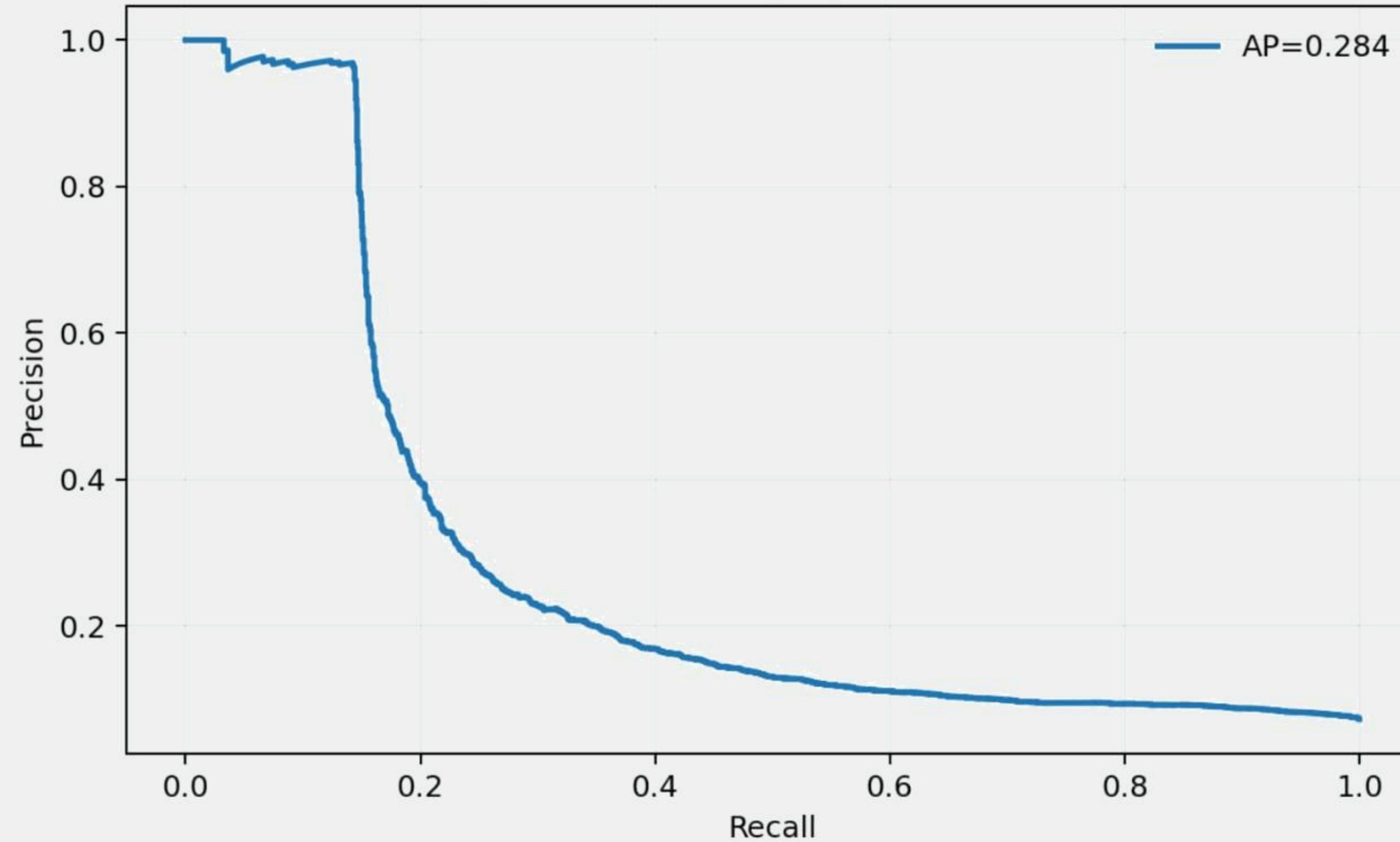
## **Shapley Flow: Создание До 1800 (старые мастера) (n=5396)**



## AIC visibility (CLEAN v2, LogReg): ROC (val)



## AIC visibility (CLEAN v2, LogReg): Precision-Recall (val)



## AIC visibility (CLEAN v2, LogReg): Calibration (val)

