Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Искусственый интеллект(Машинное обучение)»

Студент: А. О. Дубинин
Преподаватель:
Группа: М8О-306Б
Дата:
Оценка:
Подпись:

Москва, 2020

Лабораторная работа №1

Задача: Необходимо сформировать два набора данных для приложений машинного обучения. Первый датасет должен представлять из себя табличный набор данных для задачи классификации. Второй датасет должен быть отличен от первого, и может представлять из себя набор изображений, корпус документов, другой табличный датасет или датасет из соревнования Kaggle, предназначенный для решения интересующей вас задачи машинного обучения. Необходимо провести анализ обоих наборов данных, поставить решаемую вами задачу, определить признаки необходимые для решения задачи, в случае необходимости заняться генерацией новых признаков, устранением проблем в данных, визуализировать распределение и зависимость целевого признака от выбранных признаков. В отчете описать все проблемы, с которыми вы столкнулись и выбранные подходы к их решению.

Датасеты:

- TMDB 5000 Movie Dataset
- New York City Airbnb Open Data

1 New York City Airbnb Open Data

Описание входных данных
File (AB_NYC_2019.csv):

- id – id аппартаментов.
- name – название аппартаментов.
- host ID – id хозяина.
- host_name – имя хозяина.
- neighbourhood_group – район нью йорка.
- neighbourhood – окрестность.
- latitude – широта.
- longitude – долгота.
- room_type – тип аппартаментов.
- price – цена аппартаментов.
- minimum_nights – минимальное кол-во ночей.
- number_of_reviews – кол-во отзывов.
- last_review – дата последнего отзыва.
- reviews_per_month – кол-во отзывов за месяц.
- calculated_host_listings_count – кол-во записей на аппартаменты.
- availability_365 – количество дней, когда аппартаменты доступно для бронирования.

Типы признаков

- Категориальные
  - name
  - host_name
  - neighbourhood_group

  - neighbourhood
  - room_type
  - last_review
- Количественные
  - id
  - host_id
  - latitude
  - longitude
  - price
  - number_of_reviews
  - reviews_per_month
  - calculated_host_listings_count
  - availability_365

Размер

- Строк: 48895
- Столбцов: 16

Решаемая задача

Задача предсказания: предсказать цену аппартаментов.

Будем предсказывать price.

Оставим признаки, которые понадобятся для нашей задачи

Оставим только эти колонки:

neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count, availability_365

Заполнение пропусков

	Total	Percent
reviews_per_month	10052	0.205583
rt_2	0	0.000000
availability_365	0	0.000000
longitude	0	0.000000
price	0	0.000000

reviews_per_month заполним средним значением.

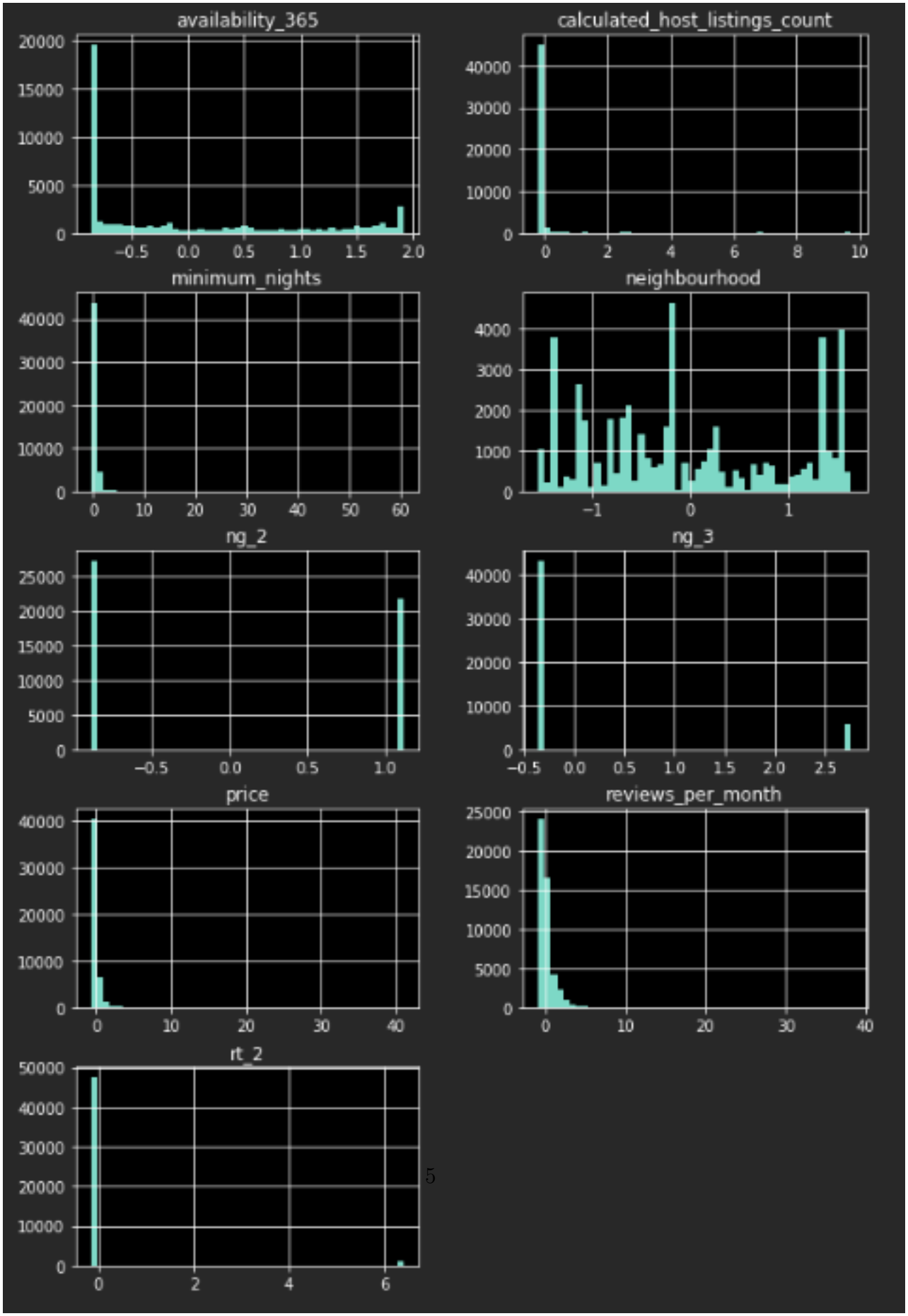
Работа с категориальными признаками

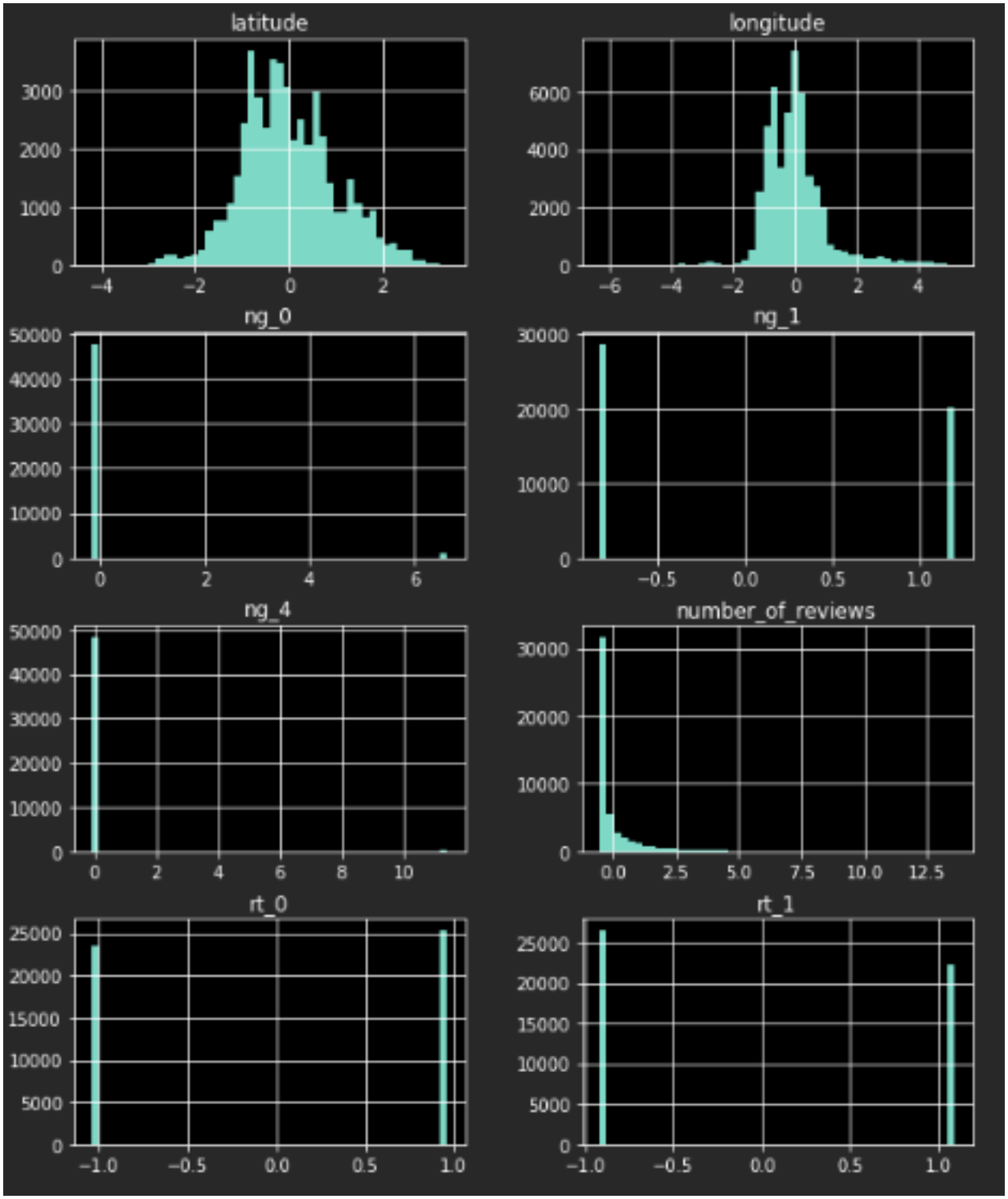
	neighbourhood_group	neighbourhood	room_type
count	48895	48895	48895
unique	5	221	3
top	Manhattan	Williamsburg	Entire home/apt
freq	21661	3920	25409

room_type и neighbourhood_group:
Кодируем one_hot_encoder'ом.

neighbourhood:
Кодируем label encoder'ом.

Распределение признаков после нормировки

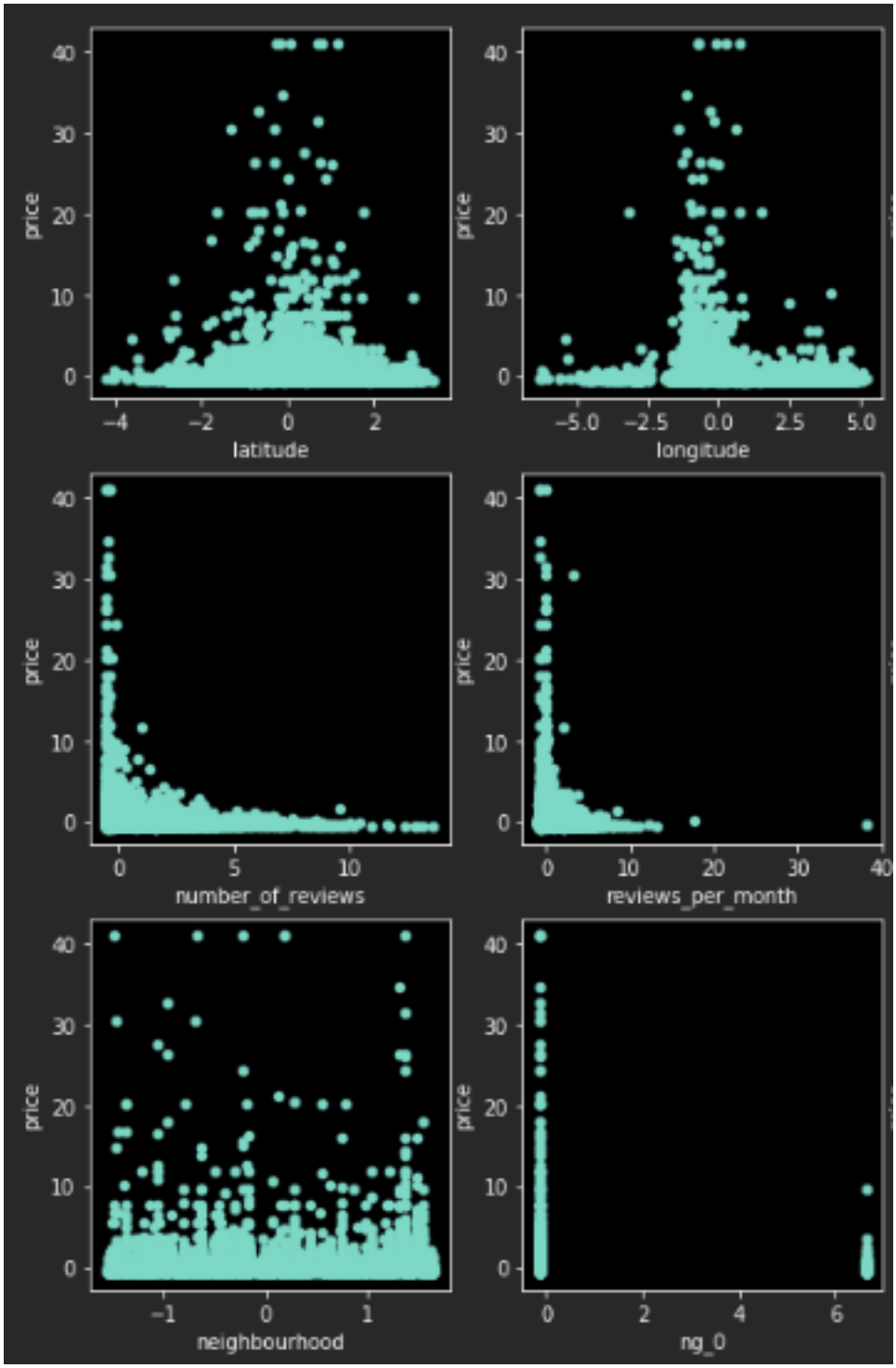

availability_365
calculated_host_listings_count
minimum_nights
neighbourhood
ng_2
ng_3
price
reviews_per_month
rt_2

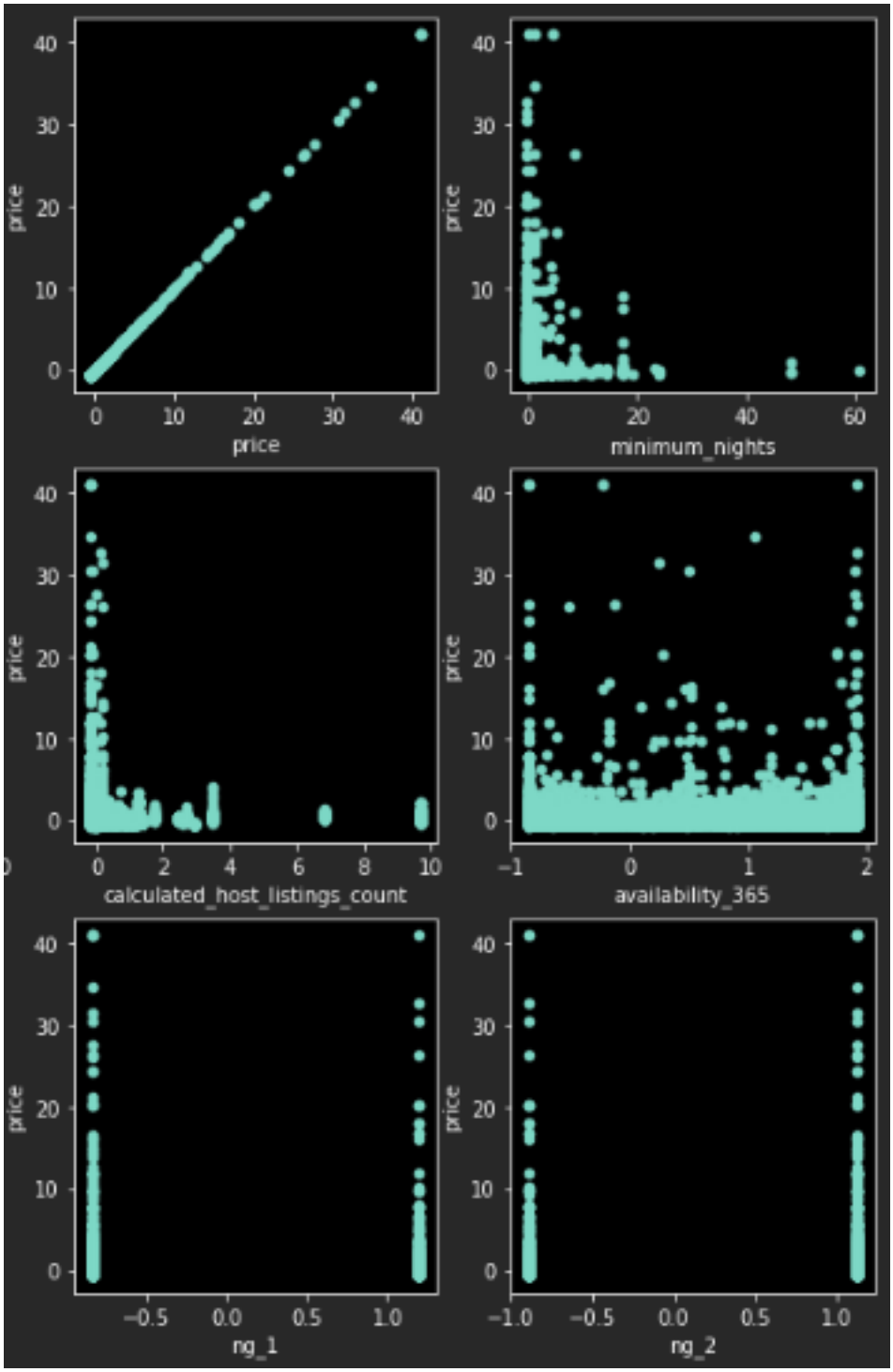


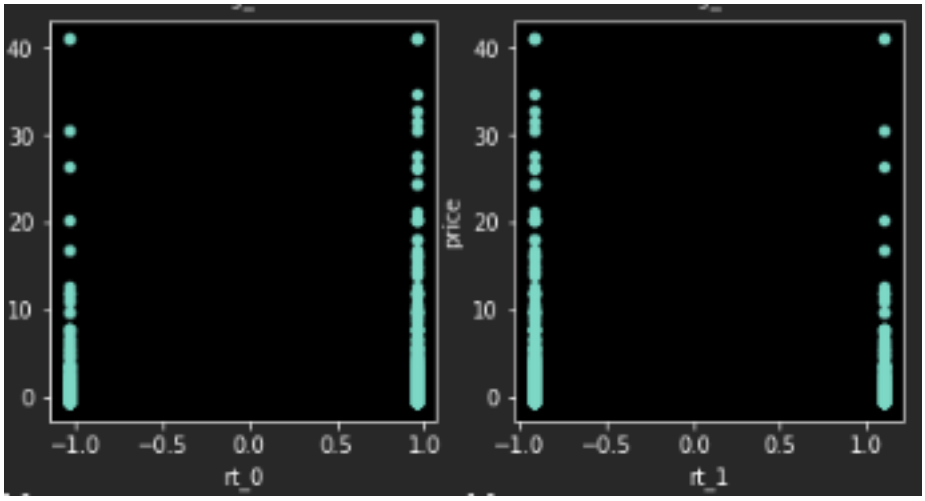
Корреляция

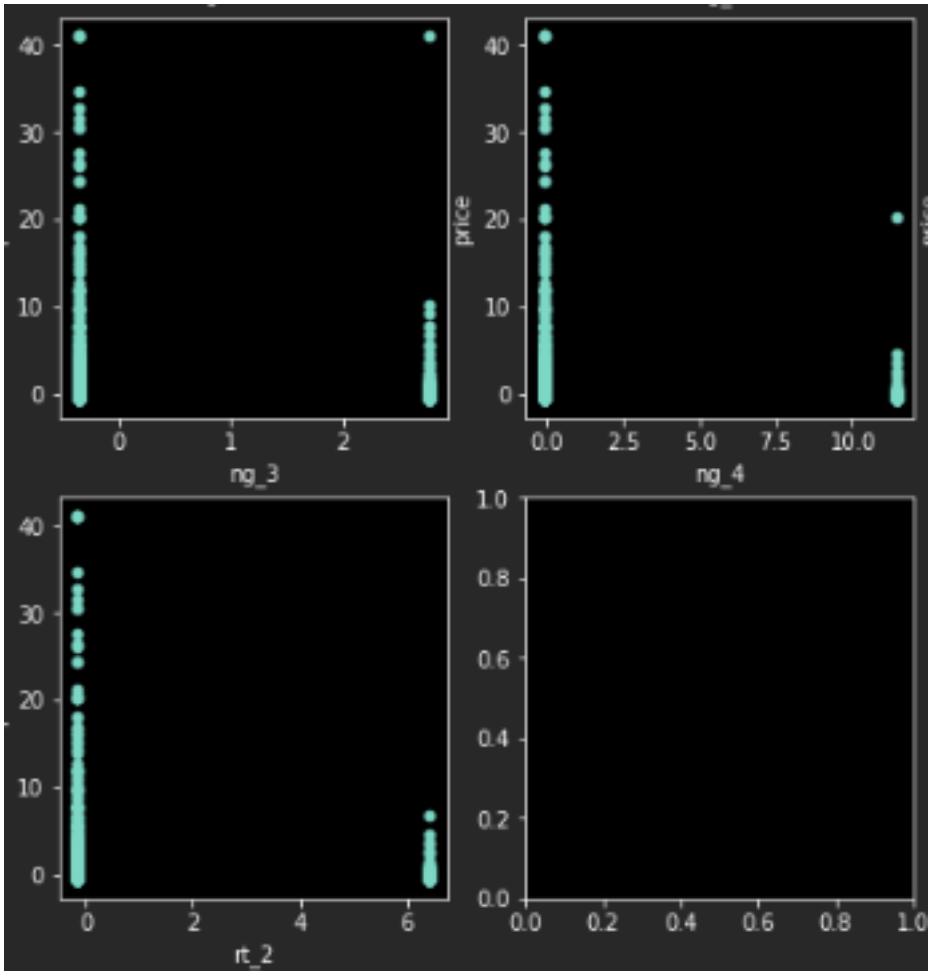
```
price                              1.000000
rt_0                               0.255857
ng_2                               0.163976
availability_365                   0.081829
neighbourhood                      0.062057
calculated_host_listings_count     0.057472
minimum_nights                     0.042799
latitude                           0.033939
ng_4                              -0.013840
reviews_per_month                 -0.022373
ng_0                              -0.041030
number_of_reviews                 -0.047954
rt_2                              -0.053613
ng_3                              -0.080205
ng_1                              -0.098603
longitude                         -0.150019
rt_1                              -0.240246
Name: price, dtype: float64
```

Зависимость главного значения от остальных


price
latitude
longitude
number_of_reviews
reviews_per_month
neighbourhood
ng_0


price
price
price
minimum_nights
price
calculated_host_listings_count
price
availability_365
price
ng_1
price
ng_2


price
rt_0
rt_1


price
ng_3
ng_4
rt_2

2 TMDB 5000 Movie Dataset

Описание входных данных

File №1(tmdb_5000_movies.csv):

- budget – бюджет фильма.
- genres – жанры фильмов, в формате список json object {'id': id, 'name':genre}.
- homepage – веб страница фильма.
- id – id фильма.
- keywords – ключевые слова фильмов, в формате список json object {'id': id, 'name':keyword}.
- original_language – оригинальный язык фильма.
- original_title – оригинальное название фильма.
- overview – синопсис фильма.
- popularity – популярность фильма.
- production_companies – компании - создатели фильма, в формате список json object {'id': id, 'name':company}.
- production_countries – страны - создатели фильма, в формате список json object {'id': id, 'name':company}.
- release_date – дата выхода фильма.
- revenue – доход фильма.
- runtime – продолжительность фильма в минутах.
- spoken_languages – языки на которых говорят в фильме, в формате список json object {'id': id, 'name':language}..
- status – статус фильма(вышел в прокат или нет).
- tagline – слоган фильма.
- title – название фильма.
- vote_average – средняя оценка фильма.

- vote_count – продолжительность фильма в минутах.

File №2(tmdb_5000_credits.csv):

- movie_id – id фильма.
- title – название фильма фильма.
- cast – актерский состав.
- crew – съемочная команда.

Анализ данных

Данные из двух файлов нужно переделать в один датасет. Из даты выхода фильма взять только год. Из стран создателей фильма взять только первую страну. Возьмем первый язык из колонки с языками, прозвучавшими в фильме. Из команды создателей фильма возьмем только режиссера. Из каста возьмем трех первых актеров. Из компаний создателей возьмем первые 3 компании. Жанры фильма запишем в одну строку с разделяющим символом, тоже самое сделаем с ключевыми словами.

Типы признаков

- Категориальные
  - genres
  - homepage
  - plot_keywords
  - language
  - original_title
  - overview
  - production_companies
  - production_countries
  - release_date
  - spoken_languages
  - status
  - tagline
  - movie_title
  - country

  - director_name
  - actor_1_name
  - actor_2_name
  - actor_3_name
  - companies_1
  - companies_2
  - companies_3
  - actor_1_name
- Количественные
  - budget
  - id
  - popularity
  - gross
  - duration
  - vote_average
  - num_voted_users
  - title_year

Размер

- Строк: 4803
- Столбцов: 29

Решаемая задача

Задача классификации: классификация на хороший/плохой фильм.

Добавим столбец Nice классификации по средней оценке, если оценка фильма больше или равна средней то, будем считать, что это хороший фильм и добавим 1 в строке с фильмом в столбец Nice, иначе 0.

Оставим признаки, которые понадобятся для нашей задачи

Оставим только эти колонки:

budget, genres, popularity, gross, duration, vote_average, num_voted_users, title_year, director_name, actor_1_name, actor_2_name, actor_3_name, companies_1, companies_2, companies_3, Nice

Заполнение пропусков

	Total	Percent
companies_3	2479	0.516136
companies_2	1417	0.295024
companies_1	351	0.073079
actor_3_name	63	0.013117
actor_2_name	53	0.011035
actor_1_name	43	0.008953
director_name	30	0.006246
duration	2	0.000416
title_year	1	0.000208
num_voted_users	0	0.000000
vote_average	0	0.000000
gross	0	0.000000
popularity	0	0.000000
genres	0	0.000000
budget	0	0.000000

duration и title_year заполним средними значениями. Остальные столбцы заполним значением unknown.

Работа с категориальными признаками

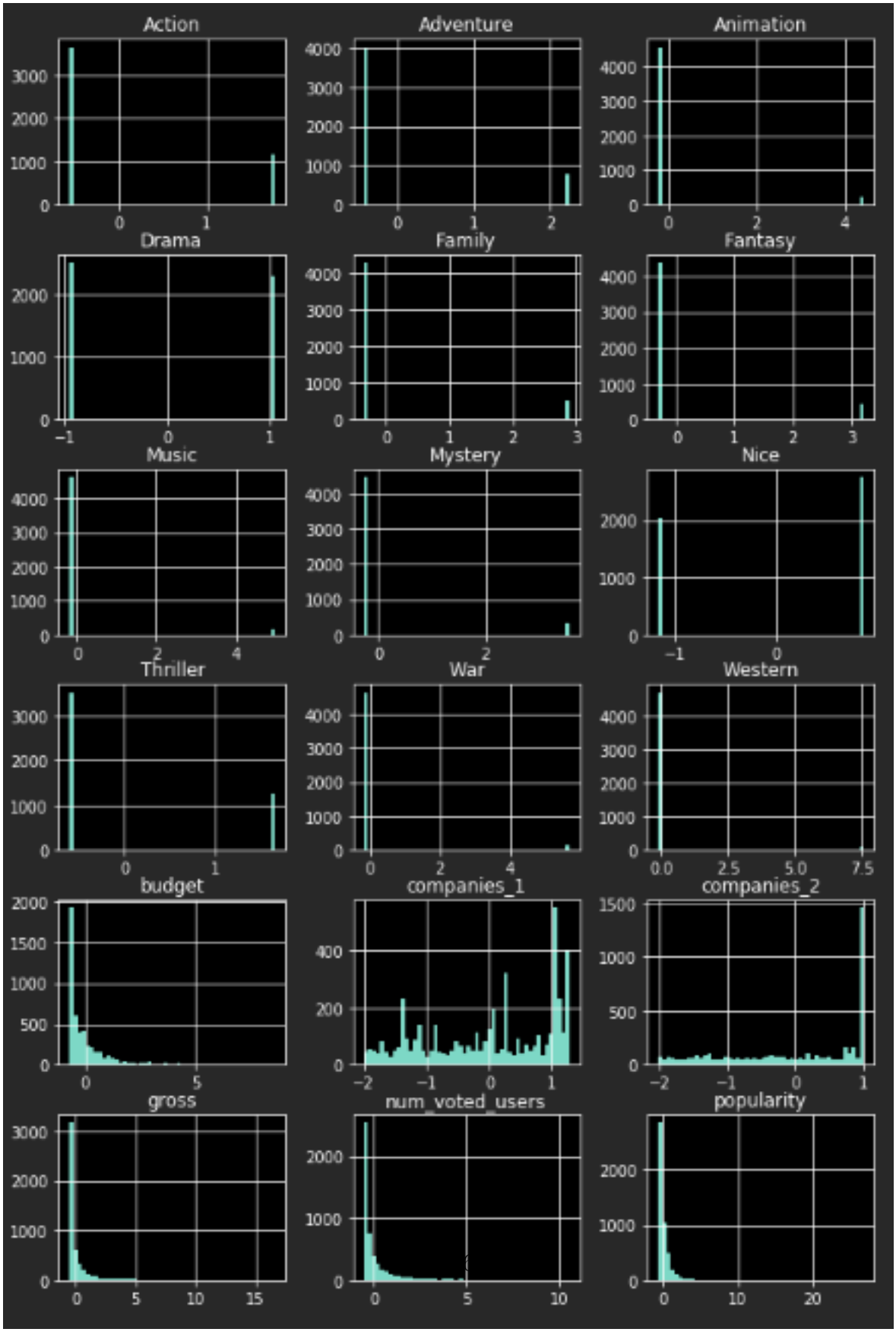
Жанры:

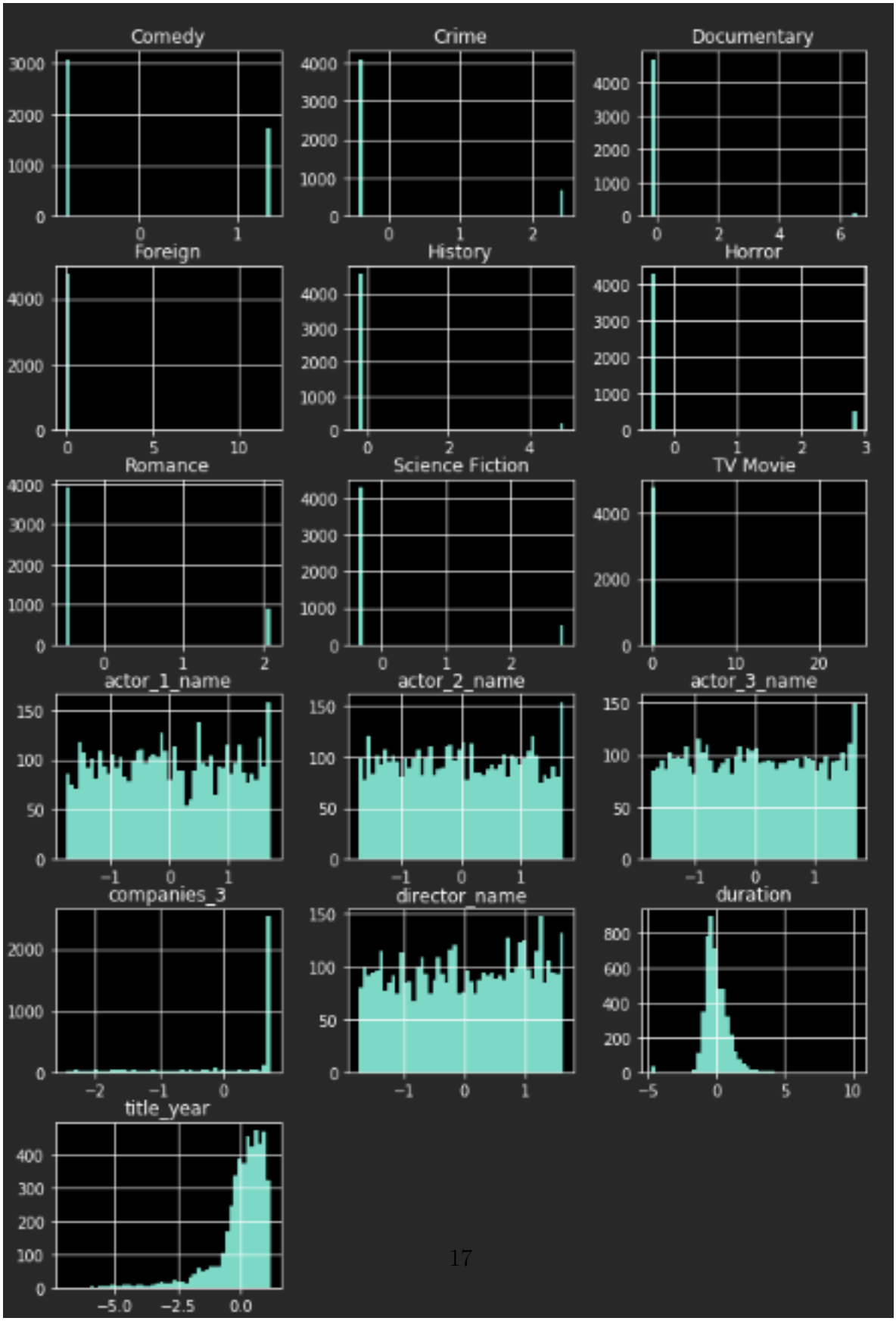
Возьмем все жанры, которые встречались в датасете. Разобьем все жанры на колонки, т.е. сделаем бинарное представление, если жанр встретился в фильме, то ставим 1 в колонке с жанром и в строке с фильмом.

Актеры, режиссеры, компании:

Кодируем label encoder'ом.

Распределение признаков после нормировки

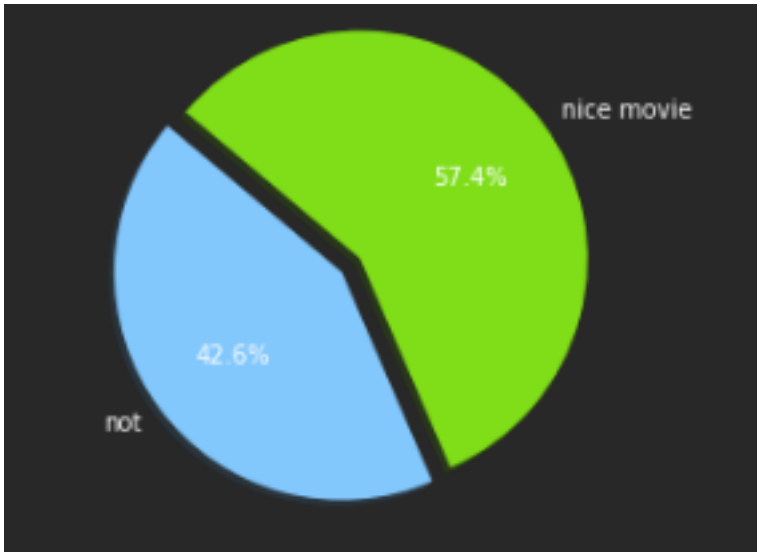

Action
Adventure
Animation
Drama
Family
Fantasy
Music
Mystery
Nice
Thriller
War
Western
budget
companies_1
companies_2
gross
num_voted_users
popularity


Comedy
Crime
Documentary
Foreign
History
Horror
Romance
Science Fiction
TV Movie
actor_1_name
actor_2_name
actor_3_name
companies_3
director_name
duration
title_year

Корреляция

```
Nice                1.000000
duration            0.304372
num_voted_users     0.252808
Drama               0.244391
popularity          0.211466
gross               0.163047
History             0.112426
War                 0.087326
Crime               0.066535
Music               0.054356
Documentary         0.050325
budget              0.035926
Mystery             0.031368
Animation           0.028805
Western             0.025831
Romance             0.023815
Foreign             0.022551
actor_3_name        0.014963
actor_1_name        0.009964
director_name       0.007537
TV Movie           -0.016421
Adventure          -0.017386
companies_3        -0.025096
actor_2_name       -0.029793
Family             -0.037299
Fantasy            -0.040506
companies_2        -0.040591
companies_1        -0.041067
Science Fiction    -0.041468
Thriller           -0.051529
Action             -0.064225
Comedy             -0.132650
Horror             -0.151625
title_year         -0.167973
Name: Nice, dtype: float64
```

Отношение классов


nice movie
57.4%
42.6%
not

3 Выводы

В ходе лабораторной работы я проанализировал 2 датасета. Посмотрев на анализ датасетов на kaggle в ноутбуках других людей понял, что для обработки признаков нужны знания статистики, например, чтобы строить различные сложные графики и уметь их понимать. Важно так же уметь создавать новые признаки из имеющихся и отбирать только самые нужные из изначальных. Так же в этой работе я научился кодировать категориальные признаки различными способами.