Задача

Разработать модель, позволяющую классифицировать музыкальные произведения по жанрам.

Обзор данных

Датасет, в котором собраны некоторые характеристики музыкальных произведений и их жанры. Признаки:

instance_id - уникальный идентификатор трека

rack_name - название трека

acousticness - акустичность

anceability - танцевальность

uration_ms -продолжительность в милисекундах

energy - энергичность

instrumentalness - инструментальность

key - тональность

iveness - привлекательность

loudness - громкость

mode - наклонение

speechiness - выразительность

tempo - темп

obtained_date - дата загрузки в сервис

valence - привлекательность произведения для пользователей сервиса

music_genre - музыкальный жанр (целевой признак)

	instance_id	track_name	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	obtained
0	25143.0	Highwayman	0.4800	0.670	182653.0	0.351	0.017600	D	0.115	-16.842	Major	0.0463	101.384	
1	26091.0	Toes Across The Floor	0.2430	0.452	187133.0	0.670	0.000051	A	0.108	-8.392	Minor	0.0352	113.071	
2	87888.0	First Person on Earth	0.2280	0.454	173448.0	0.804	0.000000	E	0.181	-5.225	Minor	0.3710	80.980	
3	77021.0	No Te Veo - Digital Single	0.0558	0.847	255987.0	0.873	0.000003	G#	0.325	-4.805	Minor	0.0804	116.007	
4	20852.0	Chasing Shadows	0.2270	0.742	195333.0	0.575	0.000002	C	0.176	-5.550	Major	0.0487	76.494	

Предобработка данных

В процессе предобработки данных по сумме ключевых параметров ('track_name', 'duration_ms', 'acousticness', 'key', 'tempo', 'danceability', 'duration_ms', 'energy', 'instrumentalness', 'mode', 'speechiness') было найдено 485 дубликатов. Дубликаты были удалены, чтобы модель не сместилась в сторону дублированного класса.

В столбце 'duration_ms' были найдены аномальные значения со знаком минус. На данном этапе заменили аномальные значения на NaN, чтобы избежать искажения результатов вычислений. Также данные столбца 'duration_ms' перевели в минуты для удобства восприятия.

Столбцы 'track_name', 'duration_ms', 'obtained_date' удалены за ненадобностью.

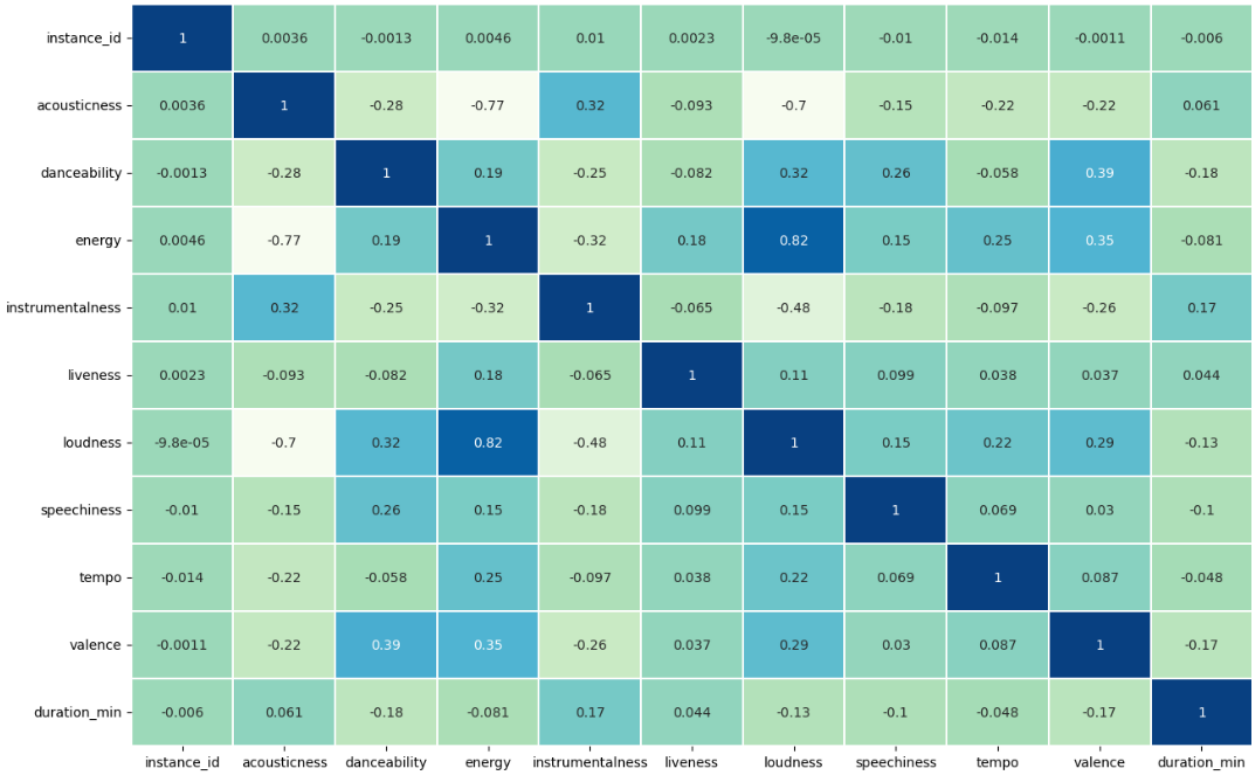
При знакомстве с данными обнаружены пропущенные значения в следующих столбцах:

- key - 735 значений,
- mode - 506 значений,
- tempo - 442 значения.

На данном этапе пропущенные значения трогать не стали.

Исследовательский анализ данных

В процессе исследовательского анализа данных изучили корреляцию между данными.



Высокая корреляция между признаками:

- «Акустичность» и «энергия» (отрицательная, -0,77)
- «Акустичность» и «громкость» (отрицательная, -0,70)
- «энергия» и «громкость» (положительная, 0.82)

Средняя корреляция между признаками:

- «Акустичность» и «инструментальность» (положительная, 0,32)
- «Танцевальность» и «привлекательность» (положительная, 0,39)
- «Энергия» и «инструментальность» (отрицательная, -0.32)
- «Энергия» и «привлекательность» (положительная, 0,35)
- «Инструментальность» и «громкость» (отрицательная, -0,48)

Высокая корреляция может говорить о мультиколлинеарности. Это может плохо влиять на модели машинного обучения. Поэтому данные были проверены на мультиколлинеарность по фактору инфляции вариаций (VIF).

	feature	VIF
0	acousticness	3.835979
1	danceability	11.717950
2	energy	16.536895
3	instrumentalness	1.796055
4	loudness	8.315741
5	valence	6.537198
6	tempo	15.071524
7	liveness	2.564957
8	speechiness	2.113525
9	duration_min	5.879428

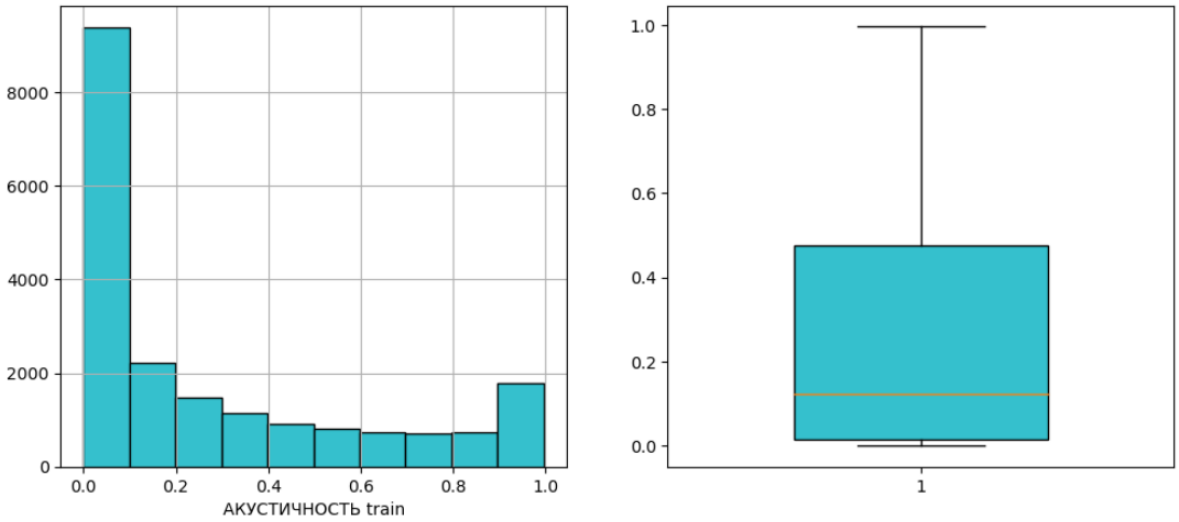
Значения более 10 говорят об очень высокой мультиколлинеарности. Высокие значения VIF у столбцов 'energy', 'tempo' и 'danceability'. В дальнейшем с этими признаками можно будет поработать.

В столбцах 'danceability', 'instrumentalness', 'liveness', 'loudness', 'speechiness', 'tempo', 'duration_min' обнаружены выбросы. Выбросы не удалены, т.к. подобные значения присутствуют и в тестовом датафрейме. Гистограммы и боксплоты для всех признаков датафрейма представлены ниже.

Описание признака АКУСТИЧНОСТЬ:

```
count    19909.000000
mean         0.276809
std          0.322891
min          0.000000
25%          0.015300
50%          0.122000
75%          0.475000
max          0.996000
Name: acousticness, dtype: float64
```

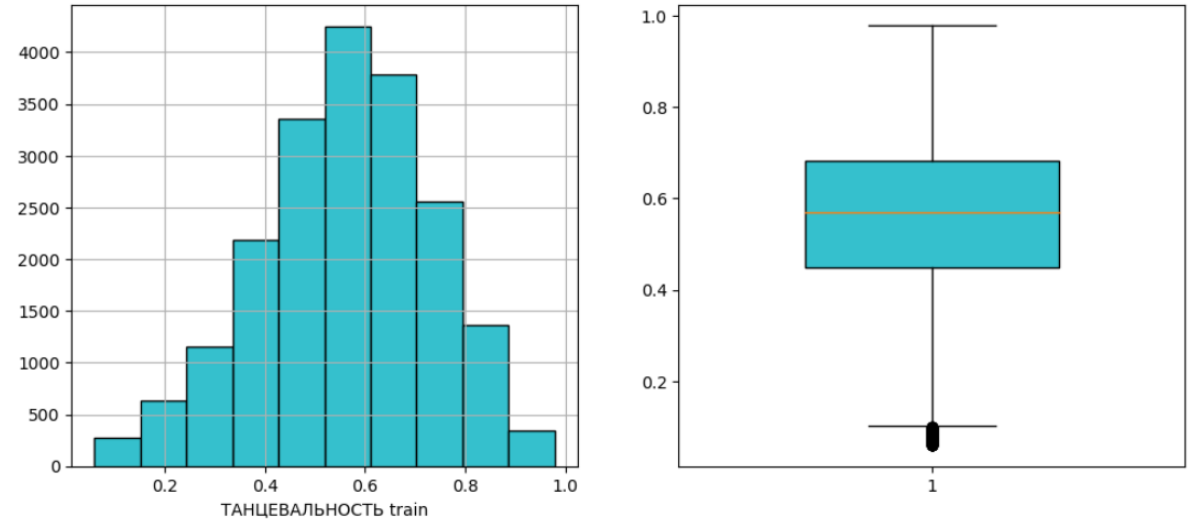
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА АКУСТИЧНОСТЬ:



Описание признака ТАНЦЕВАЛЬНОСТЬ:

```
count    19909.000000
mean         0.560969
std          0.172067
min          0.060000
25%          0.450000
50%          0.569000
75%          0.682000
max          0.978000
Name: danceability, dtype: float64
```

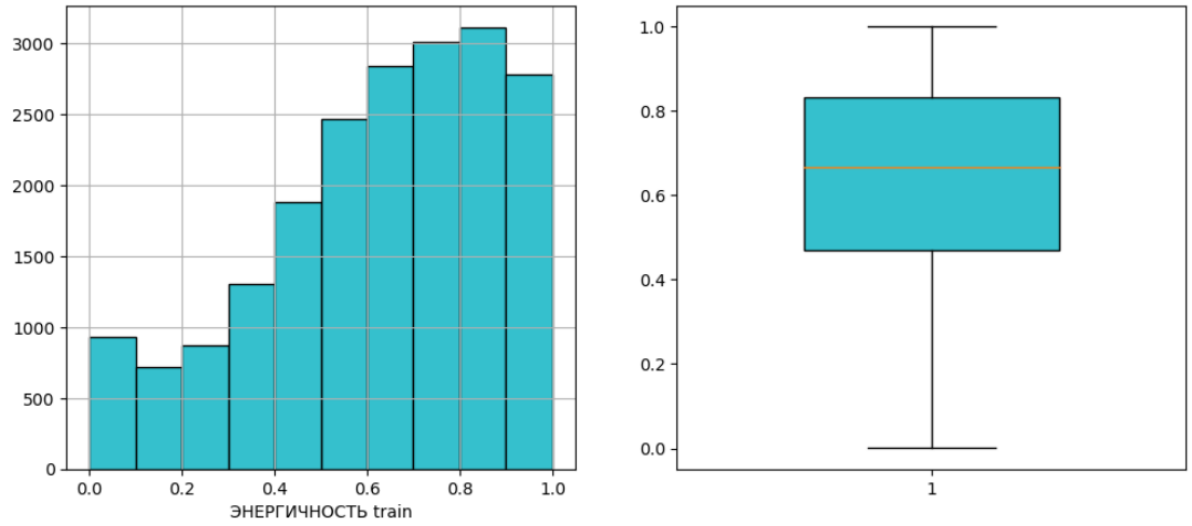
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ТАНЦЕВАЛЬНОСТЬ:



Описание признака ЭНЕРГИЧНОСТЬ:

```
count    19909.000000
mean         0.623956
std          0.252278
min          0.001010
25%          0.468000
50%          0.665000
75%          0.830000
max          0.999000
Name: energy, dtype: float64
```

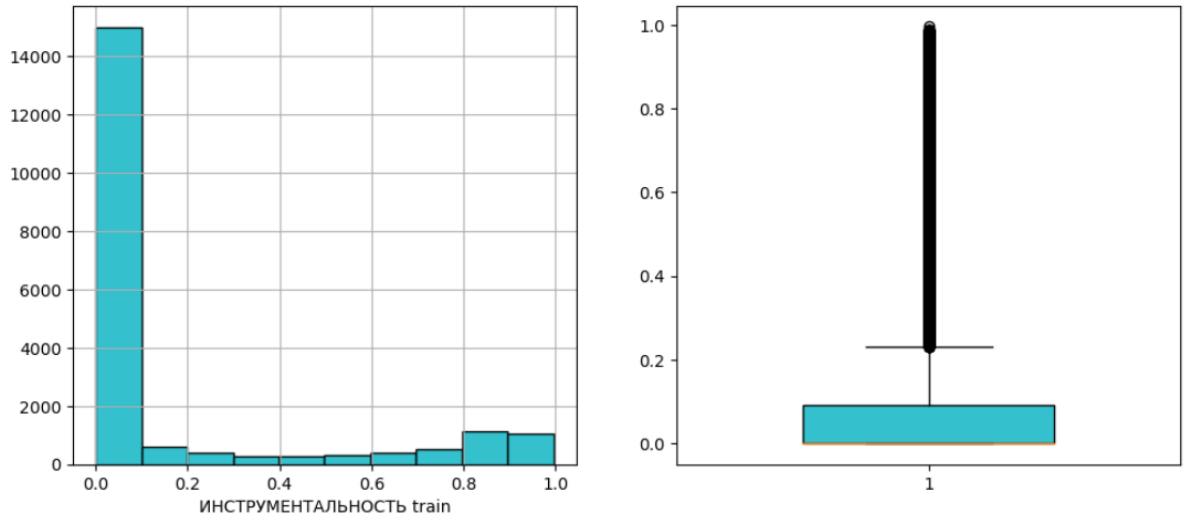
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ЭНЕРГИЧНОСТЬ:



Описание признака ИНСТРУМЕНТАЛЬНОСТЬ:

```
count    19909.000000
mean         0.162321
std          0.308292
min          0.000000
25%          0.000000
50%          0.000157
75%          0.092300
max          0.996000
Name: instrumentalness, dtype: float64
```

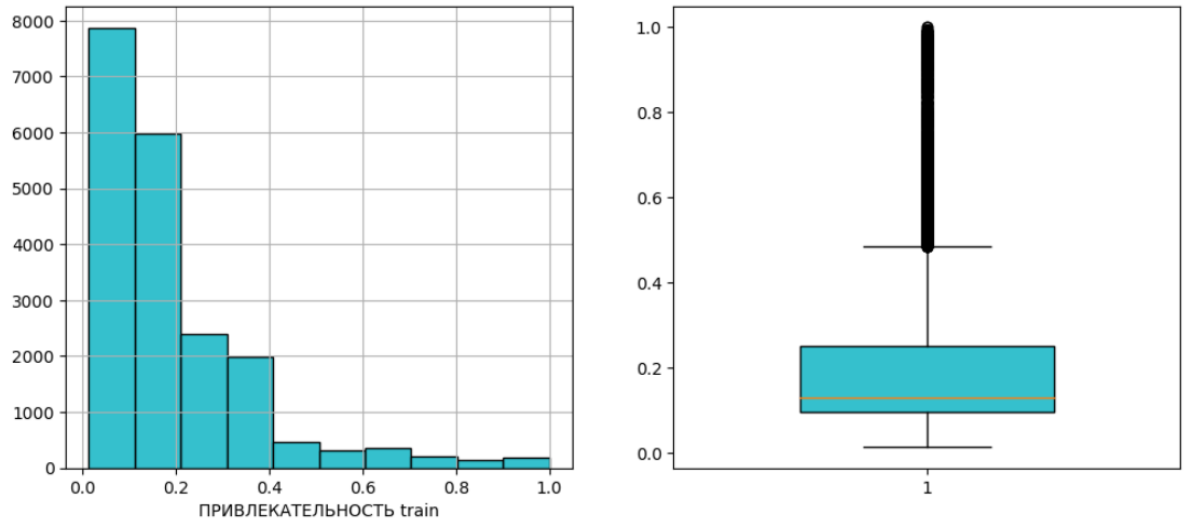
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ИНСТРУМЕНТАЛЬНОСТЬ:



Описание признака ПРИВЛЕКАТЕЛЬНОСТЬ:

```
count    19909.000000
mean         0.198568
std          0.167186
min          0.013600
25%          0.097300
50%          0.129000
75%          0.252000
max          1.000000
Name: liveness, dtype: float64
```

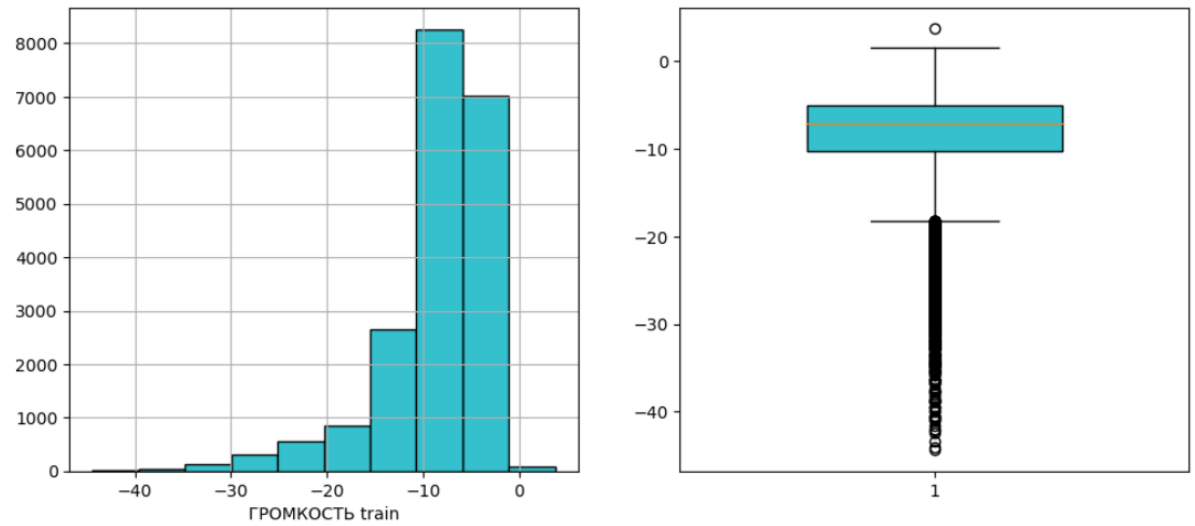
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ПРИВЛЕКАТЕЛЬНОСТЬ:



Описание признака ГРОМКОСТЬ:

```
count    19909.000000
mean        -8.592035
std          5.538925
min        -44.406000
25%        -10.315000
50%         -7.073000
75%         -5.059000
max          3.744000
Name: loudness, dtype: float64
```

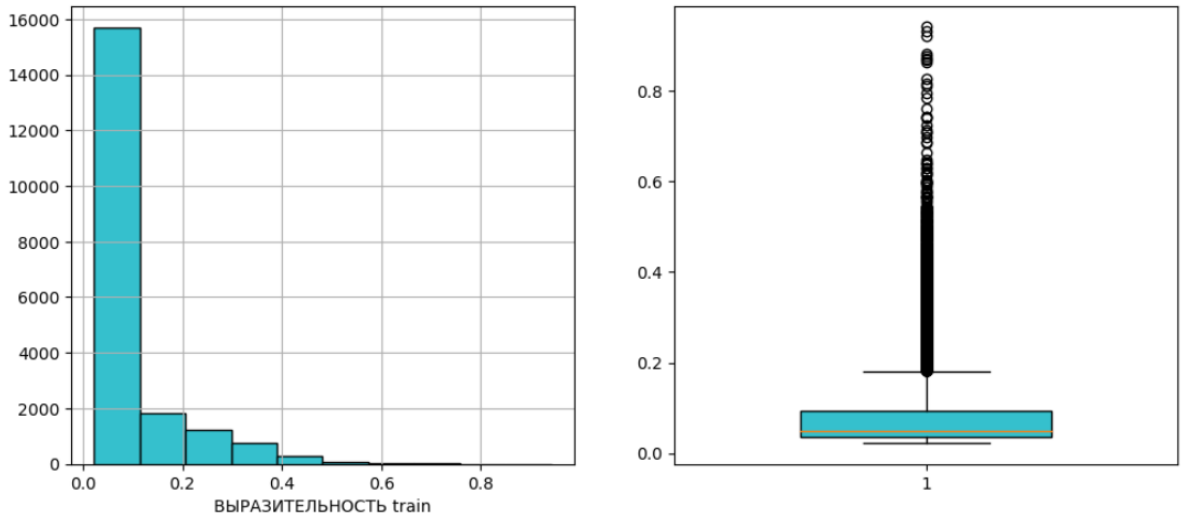
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ГРОМКОСТЬ:



Описание признака ВЫРАЗИТЕЛЬНОСТЬ:

```
count    19909.000000
mean         0.090745
std          0.097284
min          0.022300
25%          0.035600
50%          0.048900
75%          0.094100
max          0.942000
Name: speechiness, dtype: float64
```

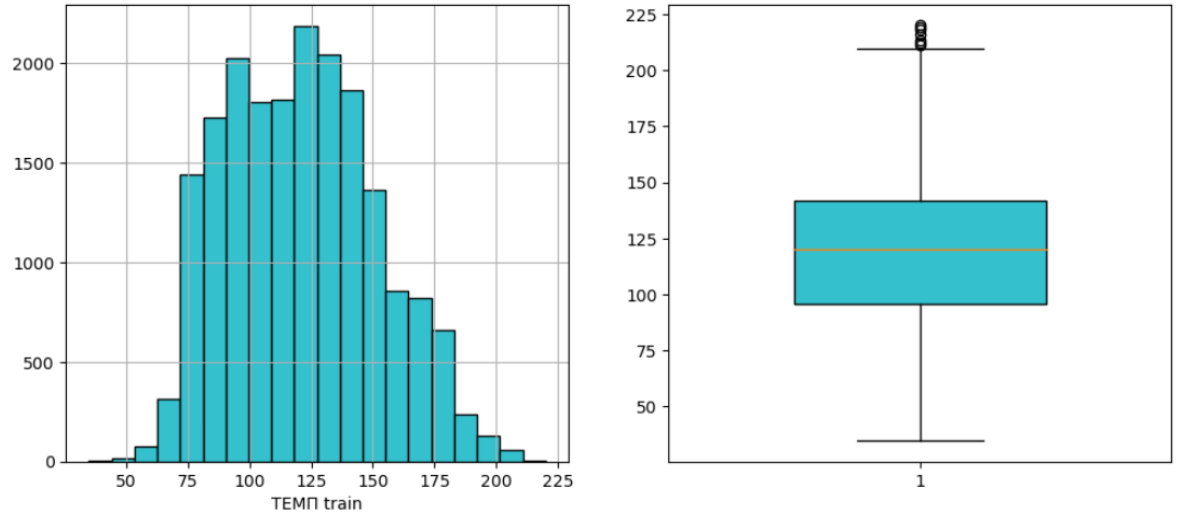
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ВЫРАЗИТЕЛЬНОСТЬ:



Описание признака ТЕМП:

```
count    19467.000000
mean       120.982157
std         30.441796
min         34.765000
25%         95.946500
50%        120.014000
75%        141.971000
max        220.041000
Name: tempo, dtype: float64
```

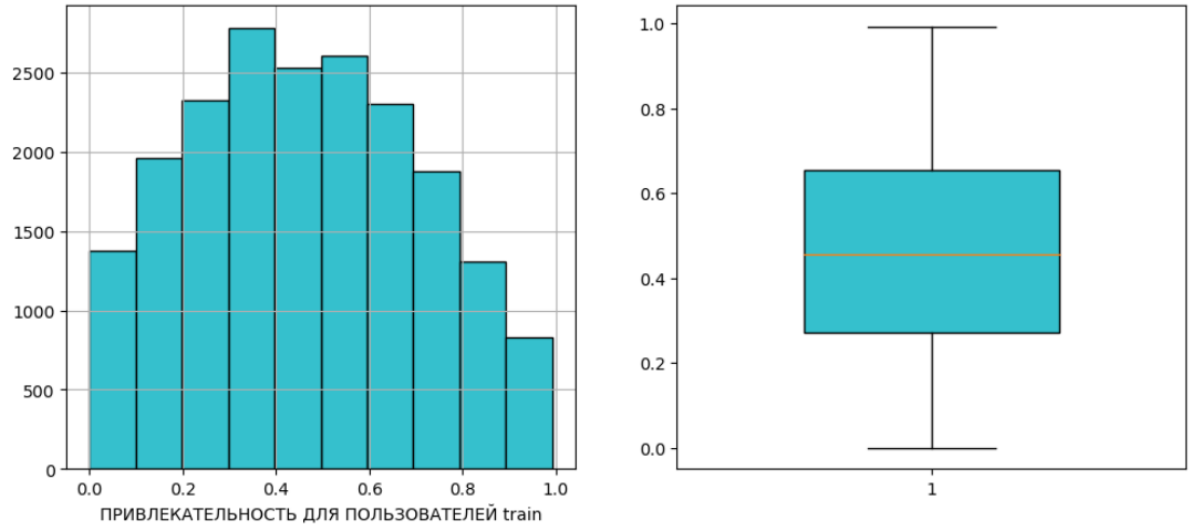
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ТЕМП:



Описание признака ПРИВЛЕКАТЕЛЬНОСТЬ ДЛЯ ПОЛЬЗОВАТЕЛЕЙ:

```
count    19909.000000
mean         0.463761
std          0.243819
min          0.000000
25%          0.271000
50%          0.456000
75%          0.653000
max          0.992000
Name: valence, dtype: float64
```

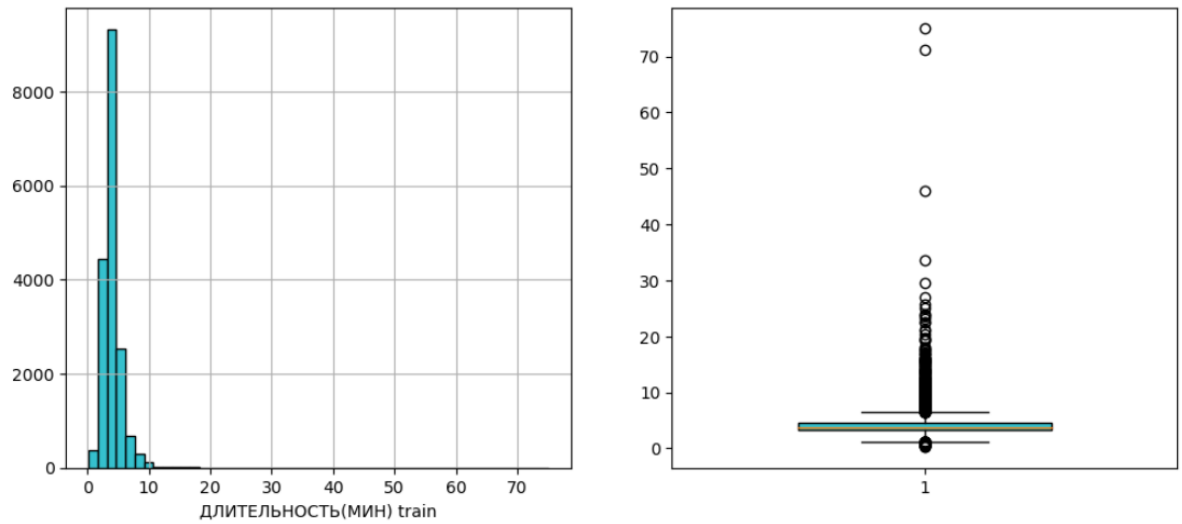
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ПРИВЛЕКАТЕЛЬНОСТЬ ДЛЯ ПОЛЬЗОВАТЕЛЕЙ:

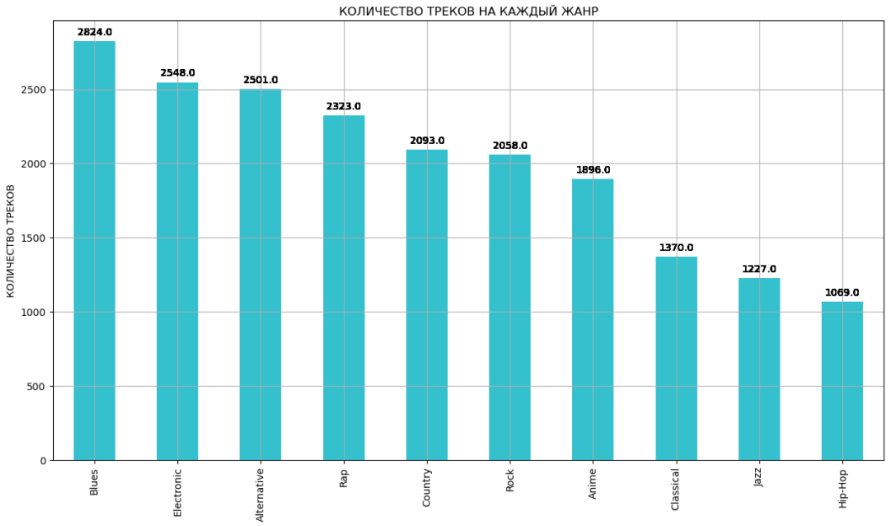


Описание признака ДЛИТЕЛЬНОСТЬ(МИН):

```
count    17907.000000
mean         4.079050
std          1.836485
min          0.260000
25%          3.200000
50%          3.790000
75%          4.550000
max         74.970000
Name: duration_min, dtype: float64
```

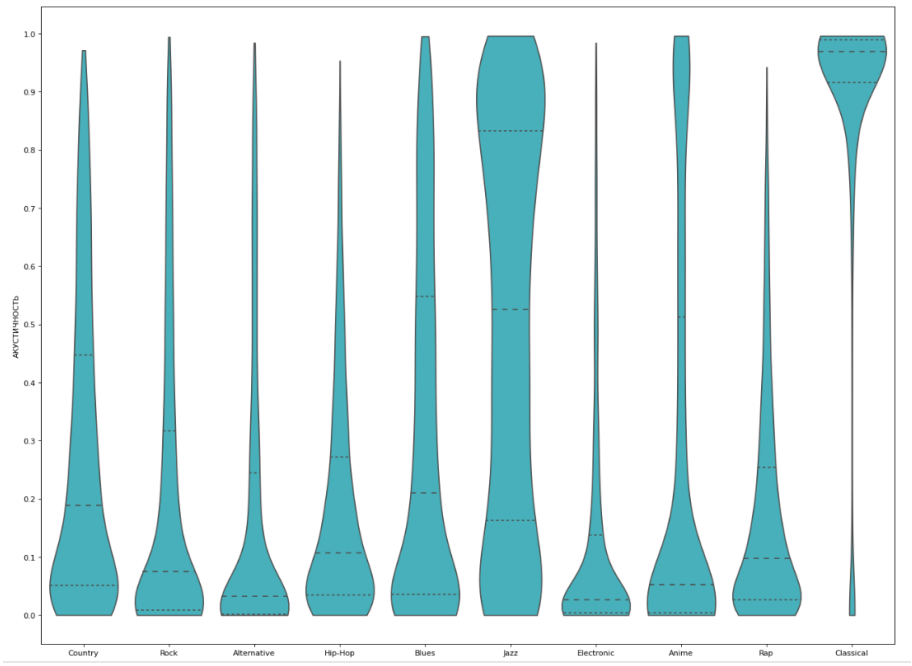
ГИСТОГРАММА И БОКСПЛОТ ДЛЯ ПРИЗНАКА ДЛИТЕЛЬНОСТЬ(МИН):



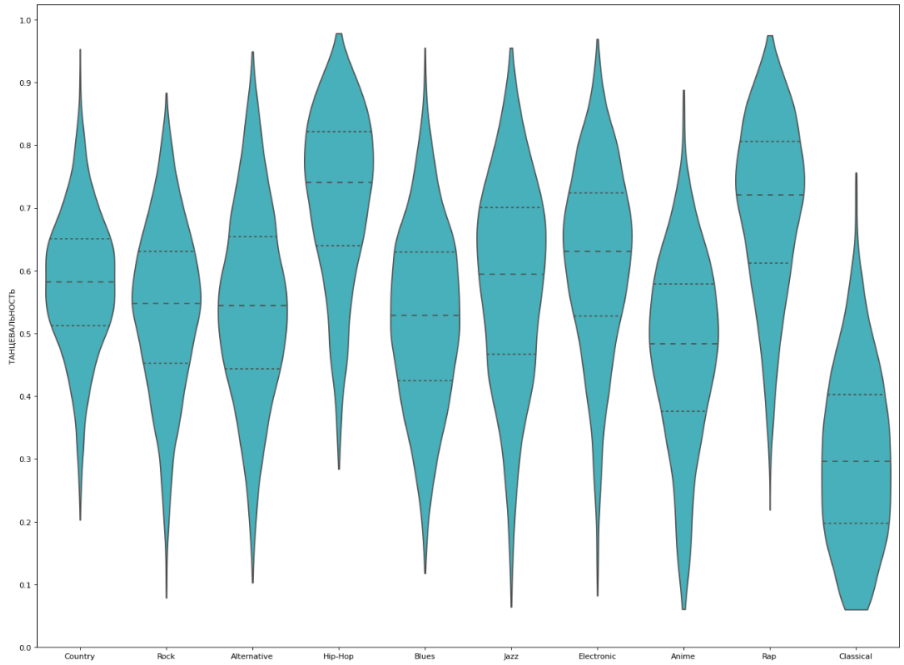


В целевом признаке наблюдается дисбаланс классов. Больше всего в наших данных блюза, а вот хип-хоп занимает последнюю строчку по количеству треков на жанр.

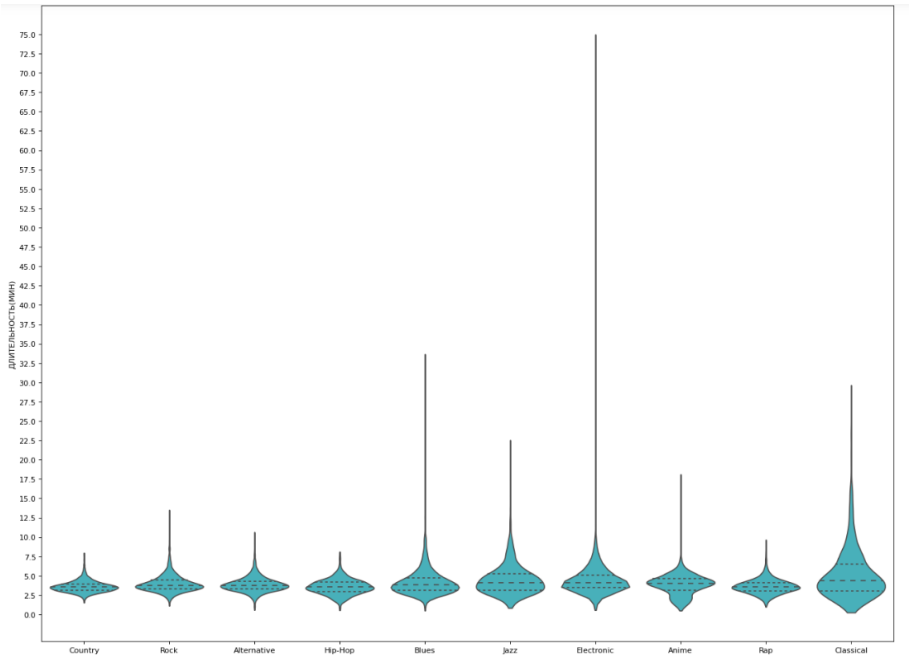
Посмотрим, как распределены значения признаков в зависимости от жанра.



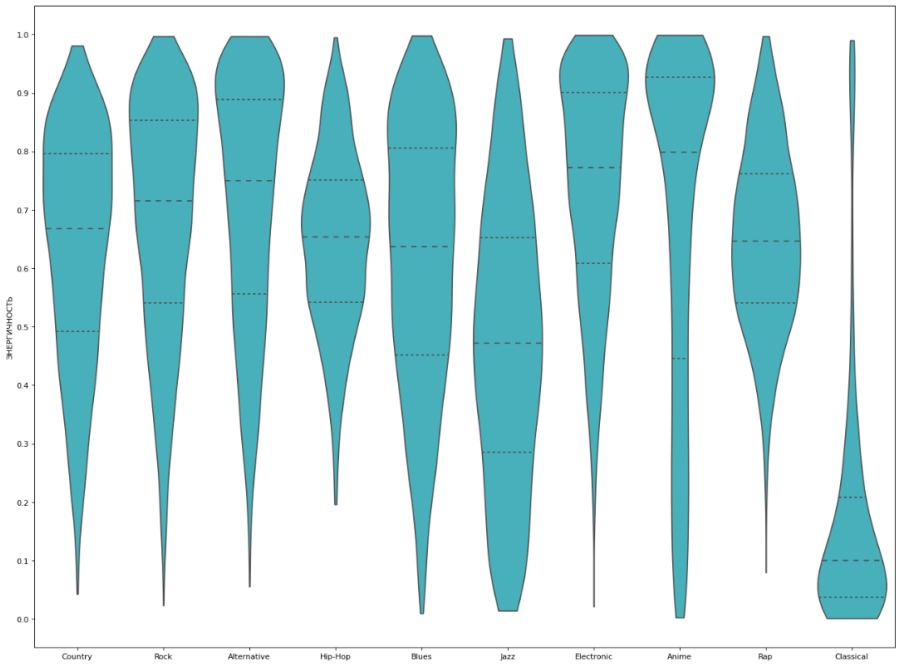
Большая часть треков в стиле **кантри, рок, альтернативная музыка, хип-хоп, блюз, электронная музыка, аниме** и **рэп** имеют акустичность примерно от 0 до 0,1. Выделяется жанр **джаз** и **классическая** музыка. **Джазовые** композиции часто встречаются с акустичностью 0-0.1, но в большей степени со значениями акустичности около 0.85-0.95. Также не малая доля треков лежит между этими двумя значениями. **Классическая** музыка сильно отличается от других жанров. Здесь большая часть треков со значениями от 0,9 до 1. Также стоит отметить, что если значение акустичности больше 0.984, то вероятнее всего это будет **джаз** или **классический** трек, менее возможно **аниме**, **блюз** или **рок**. Если значение акустичности близко к 0, то менее вероятно наткнуться на классику.



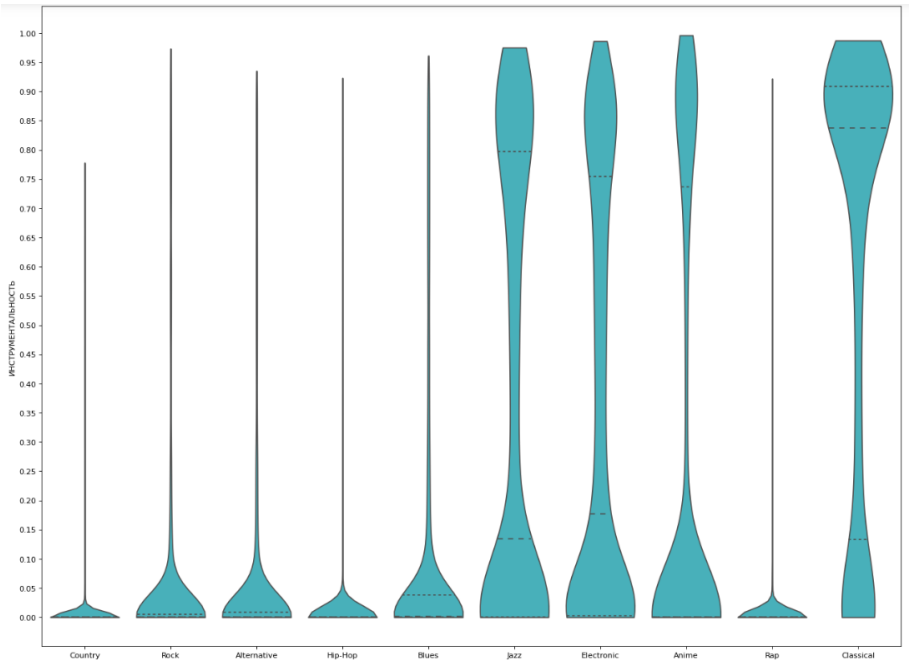
Большая часть треков в стиле кантри, рок, альтернативная музыка, блюз, джаз, электронная музыка и аниме имеют значения танцевальности от 0,45 до 0,65. Большая часть треков в стиле хип-хоп и рэп имеют значения танцевальности от 0,65 до 0,85. Классическая музыка снова выделяется: большая часть треков со значениями танцевальности от 0,20 до 0,40. Также можно отметить, что если танцевальность ниже 0.28, то это точно не хип-хоп, а если ниже 0,20 - точно не рэп и не кантри. Если значение танцевальности выше, чем 0.89, то это точно не рок и не аниме.



Большая часть треков в датафрейме длится 3-4.5 минут. Если длина трека выше 33 минут, то этот трек точно попадает в жанр электро. В жанрах кантри, хип-хоп и рэп продолжительность треков не превышает 10 минут. Треки менее 0.5 минут - это всегда классика.

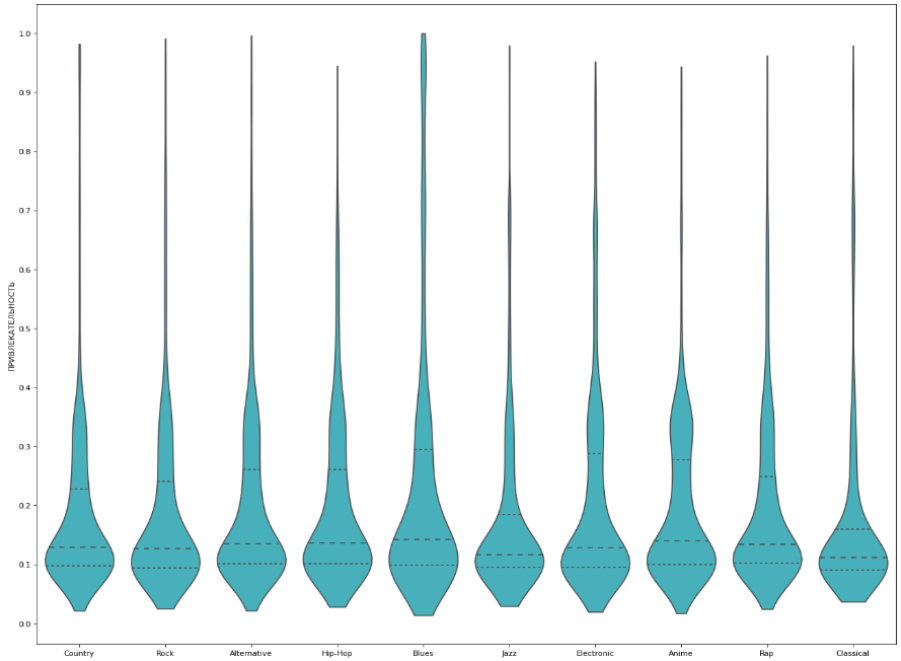


Сразу бросается в глаза, как обычно, жанр классика. В основном классические треки менее энергичны, чем треки других жанров. Значения энергичности менее 0,010 говорят о том, что трек попадает в жанр классика или аниме. Если значение энергичности менее 0,190 - это точно не хип-хоп. На максимальных значениях всё довольно тонко. Если значение энергичности более 0.981, то треки жанра кантри точно исключены. Если значение энергичности более 0.998, то трек будет из жанра аниме или электроника. Таким образом, кстати, самый большой размах получается у жанра аниме: от 0.002 до 0.999.

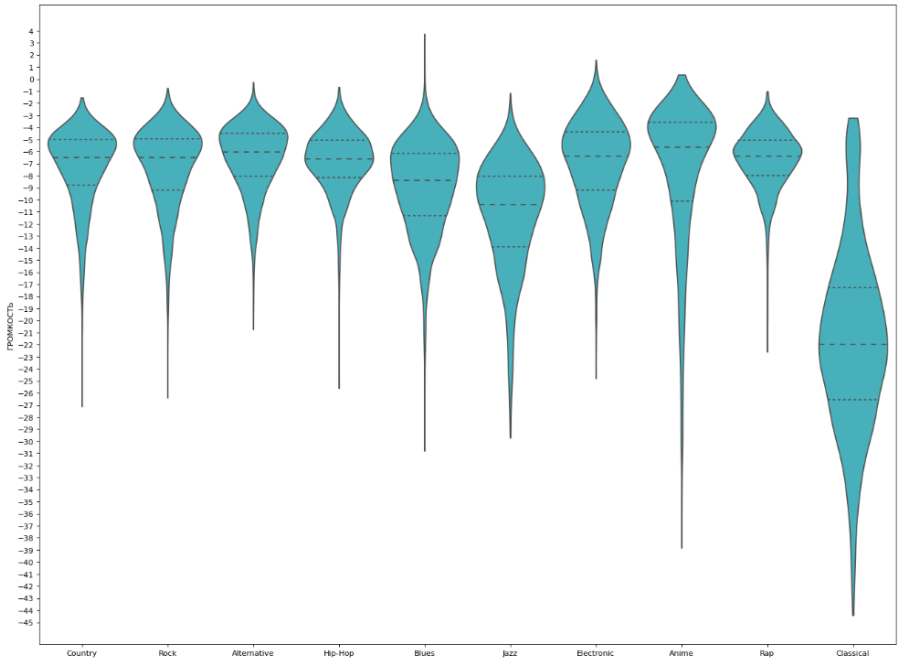


Треки жанров кантри, хип-хоп и рэп в основном имеют значения инструментальности от 0 до 0.025. Треки жанров рок, альтернатива и блюз в основном имеют значения инструментальности от 0 до 0.050. Треки жанров джаз, электроника и аниме в основном имеют значения инструментальности от 0 до 0.150. По этим жанрам также имеется меньший пик значений в диапазоне 0.750-0.95. Треки жанра классика в основном имеют значения инструментальности от 0.825 до 0.987. Меньший пик значений приходится на диапазон от 0 до

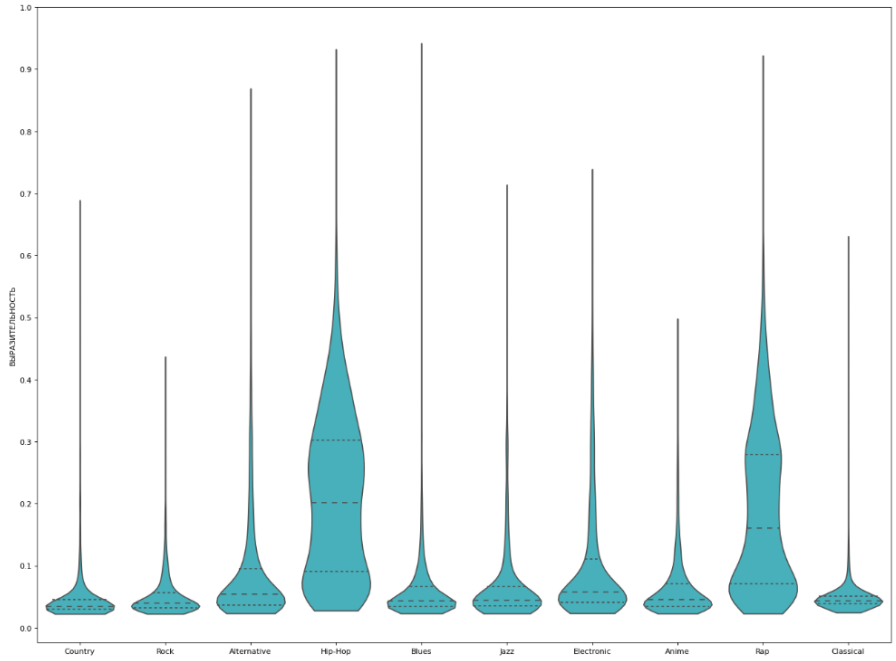
0.150. 0.120. Если значение инструментальности больше 0.990 - трек точно относится к жанру аниме.



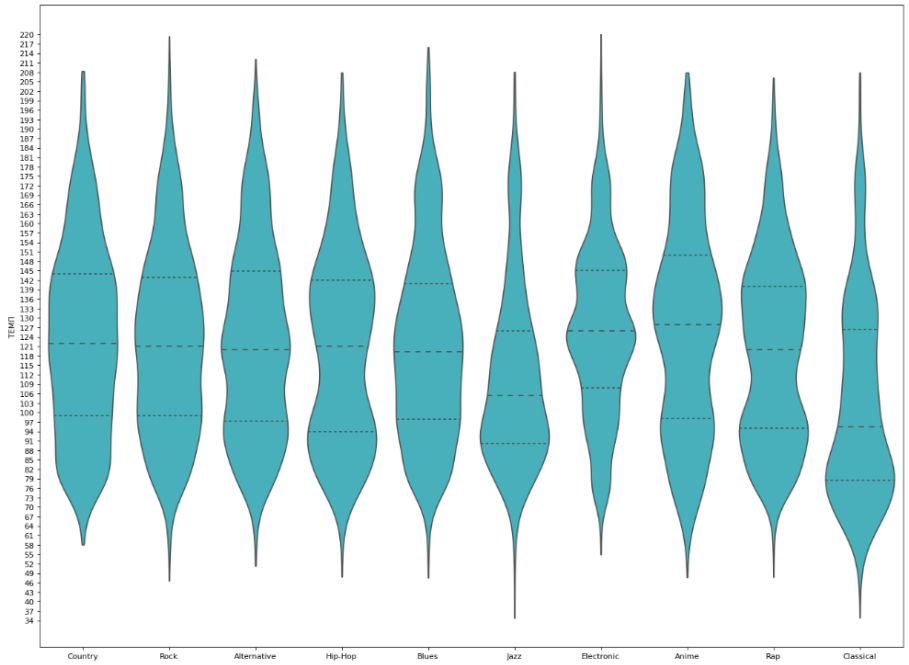
Большая часть всех треков имеет привлекательность от 0.02 до 0.20. Если привлекательность трека менее 0.036 - этот трек точно не из жанра классика. Большие значения привлекательности чаще всего у треков жанра блюз. Если значение привлекательности больше 0.996 - это точно блюз. Реже всего высокие значения привлекательности имеют жанры хип-хоп и аниме - не более 0.945.



Если значение громкости ниже -39 - то трек относится к жанру классика. Также громкость треков из этого жанра не превышает значения -3.3. диапазоне от -31 до -39 встречаются только треки жанра аниме. Громкость выше 1.5 только у блюза. Альтернативная музыка не имеет громкость ниже -20.741.

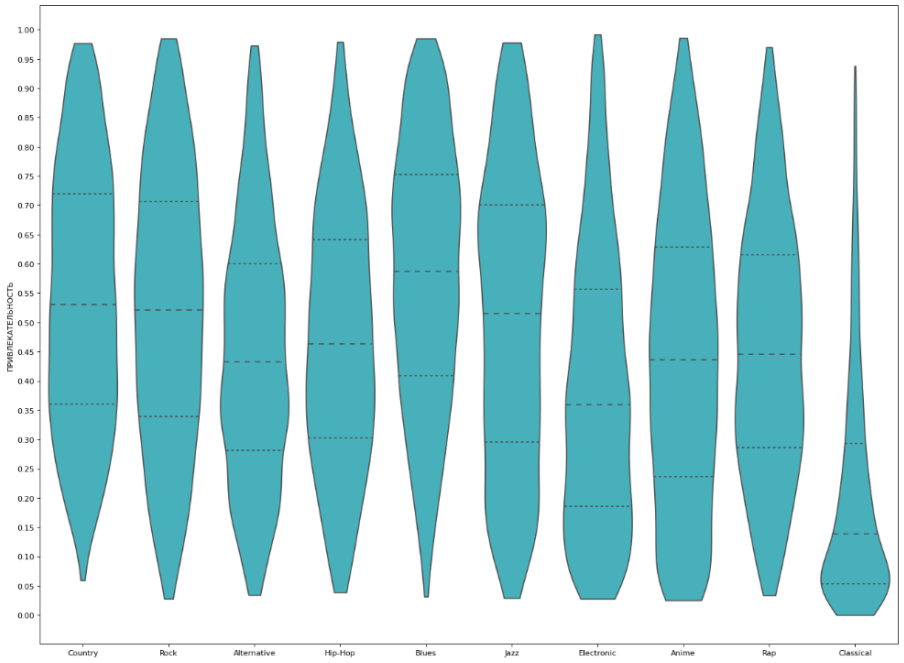


В основном выразительность всех треков варьируется от 0.02 до 0.10. Исключения составляют хип-хоп и рэп: выразительность большинства треков от 0.02 до 0.3. Самые выразительные треки встречаются в жанре блюз - диапазон выразительности от 0.932 до 0.942 принадлежит только ему. Менее выразителен рок (макс.0.437) и аниме (макс.0.498). Если выразительность трека в диапазоне 0.739 - 0.932, то трек может попадать в жанры альтернатива, хип-хоп или рэп.

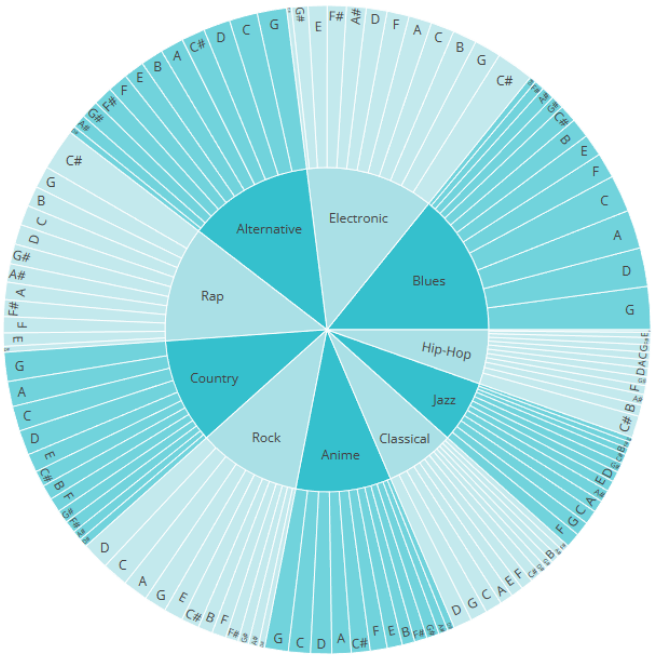


Большая часть джазовых и классических треков имеет темп ниже, чем треки из остальных жанров. В диапазон 34.765-47.587 попадают только треки из джаза и классики. Если темп ниже, чем 34.765 - то трек точно классический. Если темп ниже 58.095, то трек точно не из жанра кантри. В среднем большим темпом чаще всего отличаются треки из жанров электроника и аниме. Наивысшие значения темпа от 219.331 до 220.041 присуще

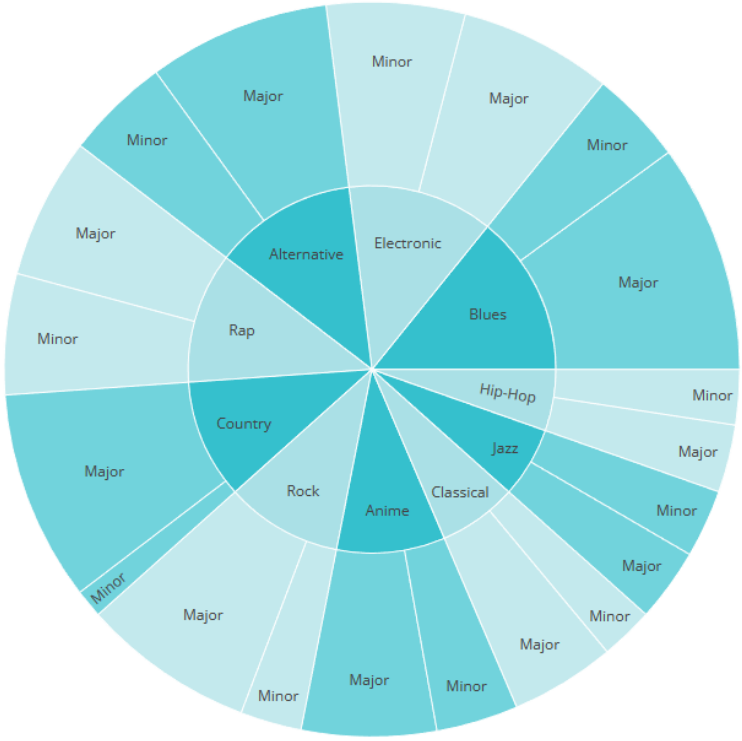
только трекам из жанра электроника. В диапазон 212.137-219.331 могут попадать также рок и блюз. Саме низкие максимальные значения у треков жанра рэп. Здесь темп никогда не превышает значений 206.247.



В среднем большая часть треков из жанра блюз для пользователей привлекательнее, чем треки остальных жанров. Наименьшей привлекательностью отличаются треки из жанра классика. Также самые низкие значения привлекательности, от 0 до 0.025 только у треков жанра классика. Если привлекательность ниже 0.059, то трек точно не из жанра кантри. Самые привлекательные треки всегда из жанра электроника: значения от 0.986 принимают только они.



Во всех жанрах, кроме классики реже всего встречается ключ D#. В классике самый редкий ключ F#.



Во всех жанрах мажор преобладает над минором. Самая большая разница в жанре кантри:1733 мажорных трека против 233 минорных.

По итогу выделим уникальные значения, которые свойственны только одному конкретному жанру.

Точно блюз:

- **Привлекательность** больше 0.996
- **Громкость** выше 1.5
- **Выразительность** 0.932-0.942

Точно электро:

- **Длина** трека выше 33 минут
- **Темп** от 219.331 и выше
- **Привлекательность** для пользователей от 0.986

Точно аниме:

- **Инструментальность** больше 0.990
- **Громкость** от -31 до -39

Точно классика:

- **Длительность** менее 0.5 минут
- **Громкость** ниже -39
- **Темп** ниже 34.765
- **Привлекательность** для пользователей 0-0.025

Поиск наилучшей модели

В данном разделе приступили к поиску наилучшей модели. За метрику взяли F-beta, со значением beta=0.5 (для большей точности).

Мы рассмотрели следующие модели:

- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier
- XGBClassifier

А также:

- Попробовали удалить, либо заменить пропущенные значения на медиану.
- Рассмотрели два способа борьбы с дисбалансом: class_weight и RandomOverSampler().
- Заменяли данные на сгруппированные.
- Удаляли мультиколлинеарные признаки.

Перебирали следующие гиперпараметры со следующими значениями:

```
params_set_LR = {
    'classifier__C': [1,3,5],
    'classifier__max_iter':[50,80,100],
    'classifier__solver': ['liblinear'],
    'classifier__penalty':['l1', 'l2'],
    'classifier__class_weight':['balanced'],
    'classifier':[LR]
}

params_set_DTC = {
    'classifier__criterion': ['gini', 'entropy'],
    'classifier__max_leaf_nodes': [8,15],
    'classifier__max_depth': [3,6,10],
    'classifier__min_samples_leaf': [8,15,20],
    'classifier__min_samples_split': [2,3,5],
    'classifier__class_weight':['balanced'],
    'classifier':[DTC]
}


params_set_RFC = {
    'classifier__n_estimators':[8,10,30],
    'classifier__max_depth': [8,12,15],
    'classifier__min_samples_split': [3,5],
    'classifier__min_samples_leaf': [3,8,12],
    'classifier__class_weight':['balanced'],
    'classifier':[RFC]
}

params_set_XGB = {
        'classifier__num_class':[10],
        'classifier__min_child_weight': [3,8],
        'classifier__gamma': [2.5,3],
        'classifier__subsample': [0.3,0.6],
        'classifier__colsample_bytree': [0.2,0.5],
        'classifier__max_depth': [3,5],
        'classifier__learning_rate': [0.1, 0.01],
        'classifier__colsample_bytree': [0.6,1.0],
        'classifier__n_estimators':[50, 250],
        'classifier':[XGB]
        }
```

Результаты при разных условиях изложены ниже:

Сравнение способов работы с пропущенными значениями: удаление и замена на медиану при следующих условиях:

- Пропущенные категориальные значения заменяем на моду.
- Способ борьбы с дисбалансом: class_weight.
- Признаки не сгуппированы.
- Признаки с высокой мультиколлинеарностью сохранены.

Удаление пропущенных значений дало чуть лучший результат для модели DTC, но замена медианой дала чуть лучший результат для модели RFC. Было принято решение заменять пропущенные значения медианой.

Модель	Удаление ПЗ	Замена медианой
LR	0.370	0.370
DTC	0.296	0.302
RFC	0.429	0.424
XGB	0.448	0.448

Замена способа борьбы с дисбалансом на RandomOverSampler при следующих условиях:

- Пропущенные численные значения заменяем на медиану.
- Пропущенные категориальные значения заменяем на моду.
- Признаки не сгуппированы.
- Признаки с высокой мультиколлинеарностью сохранены.

Модель	RandomOverSampler
LR	0.385
DTC	0.342
RFC	0.653
XGB	0.544

Группировка признаков при следующих условиях:

- Пропущенные численные значения заменяем на медиану.
- Пропущенные категориальные значения заменяем на моду.
- Способ борьбы с дисбалансом: RandomOverSampler.
- Признаки с высокой мультиколлинеарностью сохранены.

Модель	RandomOverSampler
RFC	0.554
XGB	0.356

Удаление мультиколлинеарного признака energy при следующих условиях:

- Пропущенные численные значения заменяем на медиану.
- Пропущенные категориальные значения заменяем на моду.
- Способ борьбы с дисбалансом: RandomOverSampler.

Модель	Без energy
RFC	0.645
XGB	0.535

Удаление мультиколлинеарного признака tempo при следующих условиях:

- Пропущенные численные значения заменяем на медиану.
- Пропущенные категориальные значения заменяем на моду.
- Способ борьбы с дисбалансом: RandomOverSampler.

Модель	Без tempo
RFC	0.649
XGB	0.533

В итоге, наилучший результат мы получили при следующих условиях:

- Пропущенные численные значения заменяем на **медиану**.
- Пропущенные категориальные значения заменяем на **моду**.
- Способ борьбы с дисбалансом: **RandomOverSampler**.
- Признаки **не сгуппированы**.
- Признаки с высокой мультиколлинеарностью **сохранены**.

Наилучшая модель: RandomForestClassifier

Лучшие гиперпараметры: random_state=12345, n_estimators = 30,max_depth = 20,min_samples_split = 3, min_samples_leaf = 3

Лучший результат f-score: 0.653

Применили наилучшую модель ко второй части датасета: 0.41

Полученные предсказания для тестового датафрейма выведены в отдельный файл.