Федеральное государственное образовательное бюджетное учреждение высшего образования   
**«Финансовый университет при Правительстве Российской Федерации»**

**(Финансовый университет)**

Факультет «Информационных технологий и анализа больших данных»

Департамент анализа данных машинного обучения

**Домашняя творческая работа**

по дисциплине «Технологии анализа данных и машинного обучения»

на тему:

**«Кластерный и регрессионный анализ»**

Выполнила:

студентка гр. ПИ18-3

Красотская Е. А.

Проверил:

Никитин П.В.

Москва 2020

Оглавление

[Введение 3](#_Toc61659556)

[Анализ и визуализация данных 4](#_Toc61659557)

[Кластерный анализ 12](#_Toc61659558)

[Регрессия и классификация 16](#_Toc61659559)

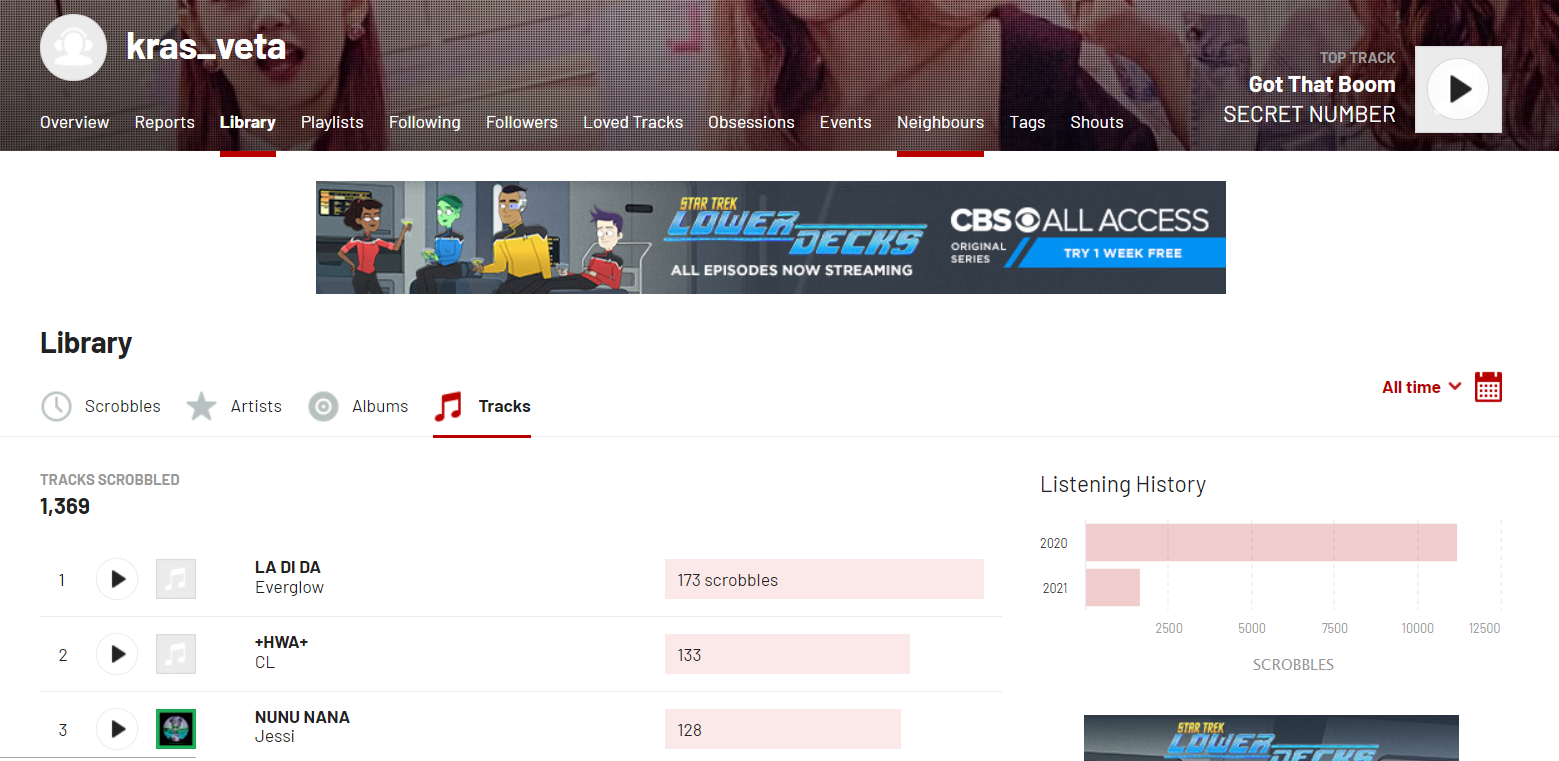
[Выводы 18](#_Toc61659560)

# **Введение**

В данной работе анализируется плейлист пользователя сайта last.fm (сайт по сбору статистики прослушиваемой на устройстве музыки). Задачей прогнозирования было поставлено предсказание заинтересованности в треке – понравится ли музыкальная композиция слушателю в будущем в зависимости от того, что он прослушал ранее.

Эта задача сама по себе заметно отличается от классического прогнозирования стоимости и т.п., поэтому почти любой результат будет удовлетворяющим для меня.

Были получены данных от трех пользователей сайта с похожими музыкальными вкусами (были взяты три пользователя, так как данных по моему профилю не было достаточно – всего 1100 записей, т.к. данный сервис только набирает популярность, таким образом, я расширила датасет).



Из библиотек были получены следующие данные о треке и исполнителе:

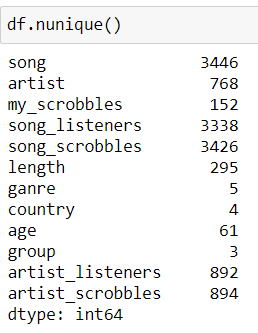
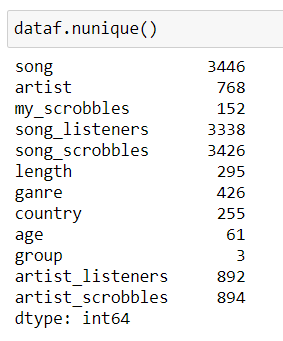
* Песня (для фильтрации возможных одинаковых строк)
* Исполнитель (для фильтрации возможных одинаковых строк)
* Мои скробблы (количество прослушиваний трека)
* Слушатели песни
* Скробблы песни
* Длина
* Жанр
* Страна и язык
* Возраст исполнителя/года карьеры группы
* Тип: группа или сольный исполнитель
* Слушатели артиста
* Скробблы артиста

Были собраны данные с 70 страниц, в сумме более 3600 треков.

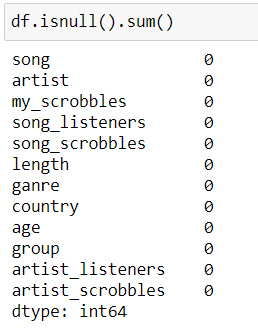
# **Анализ и визуализация данных**

До построения регрессии и кластеризации необходимо проверить данные и проанализировать их, при необходимости откорректировав.

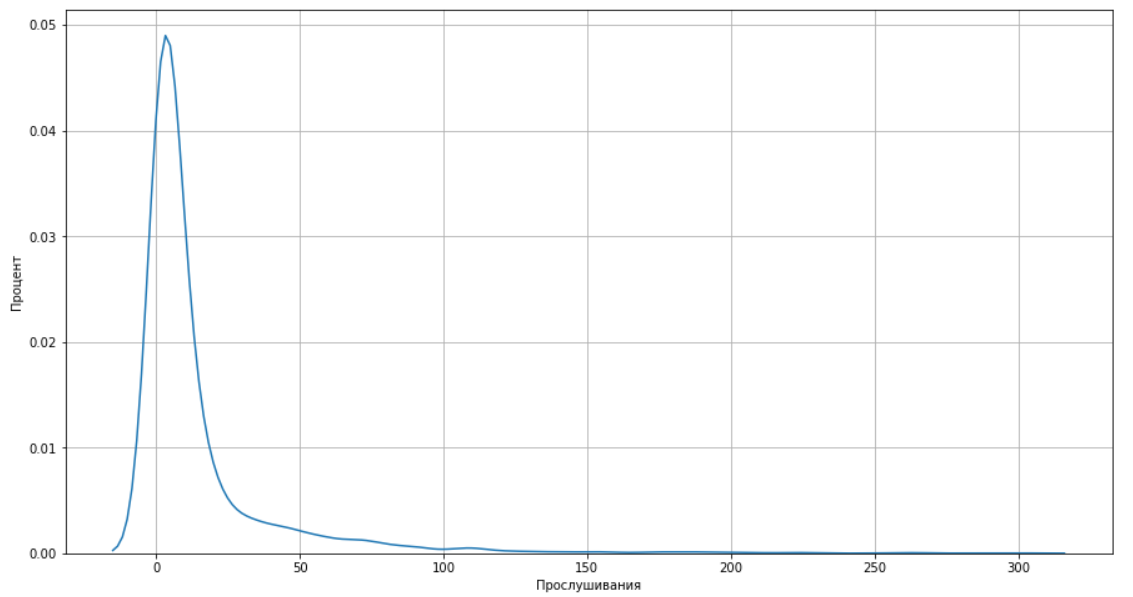
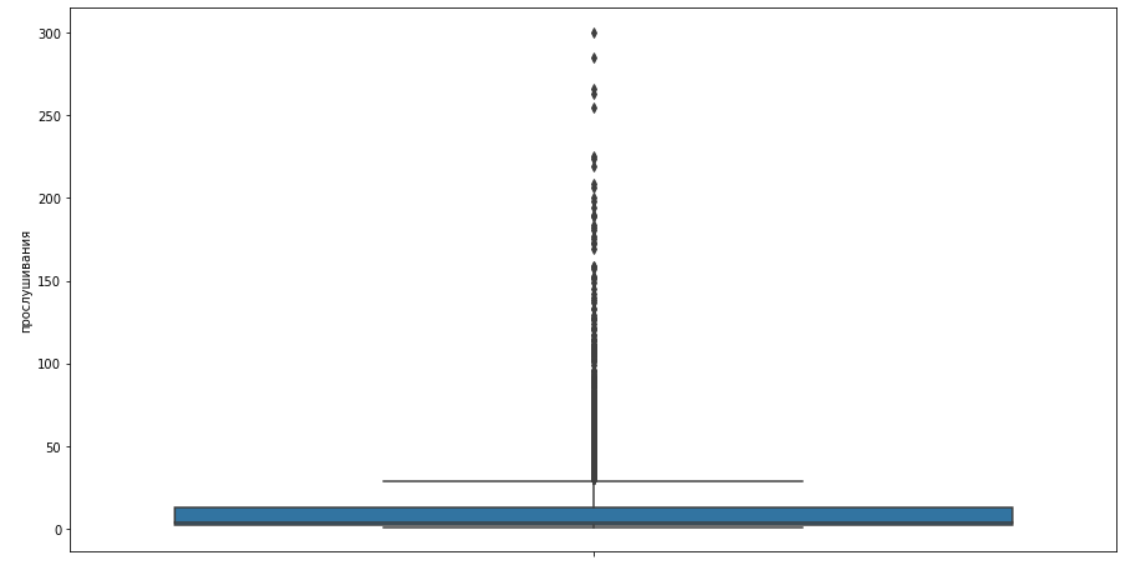
В данном случае корректировок было много: некоторые жанры, возраст, страна были пропущены в большом количестве. Некоторые данные – жанры – были классифицированы при помощи ручного анализа, а все остальные пропуски были заполнены относительно жанровой принадлежности и других данных. На скриншотах представлена работа над жанрами и странами.



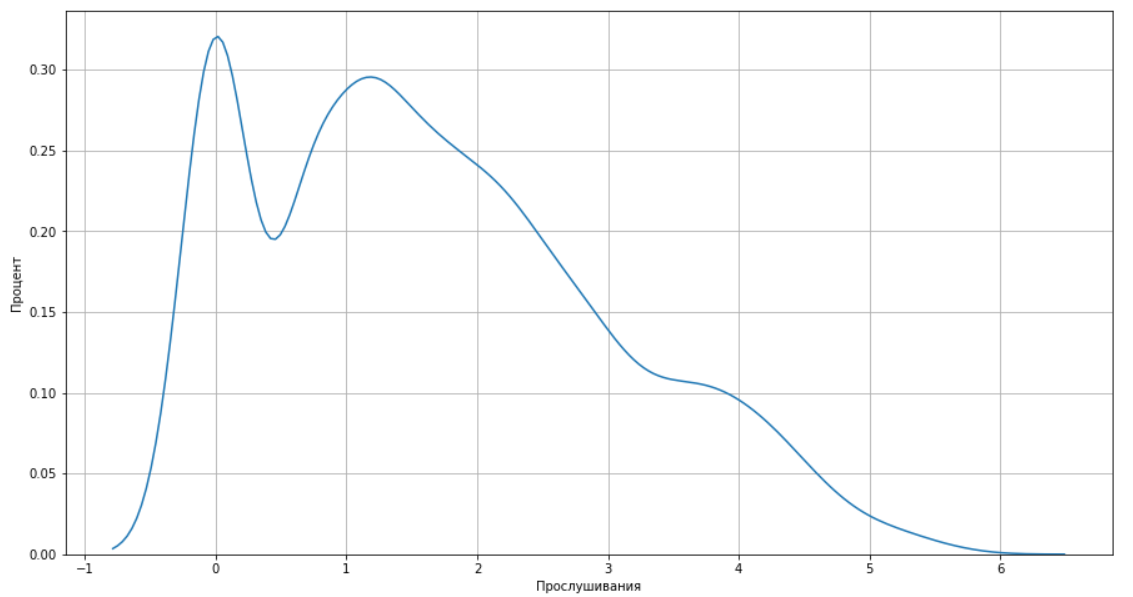
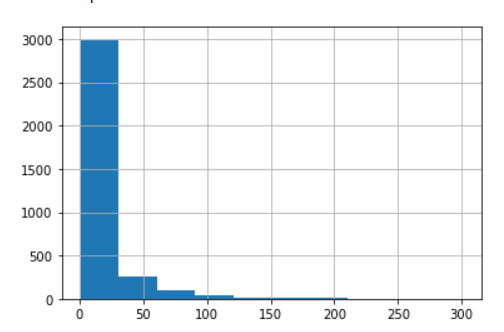
Пропуски по возрасту были заполнены с различным подходом в зависимости от типа – группа или соло-исполнитель, все пропуски ликвидированы:



Описательная статистика в основном велась в разрезе зависимости различных факторов от скробблов пользователя (прослушиваний одного трека).

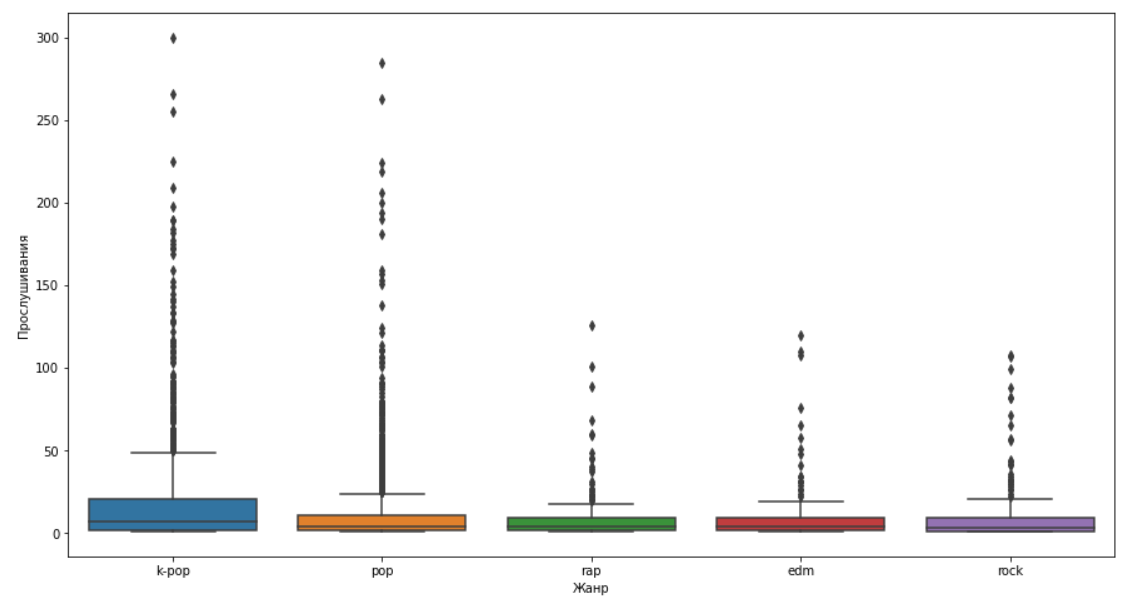


Распределение очень сильно смещено. Было использовано логарифмическое представление:

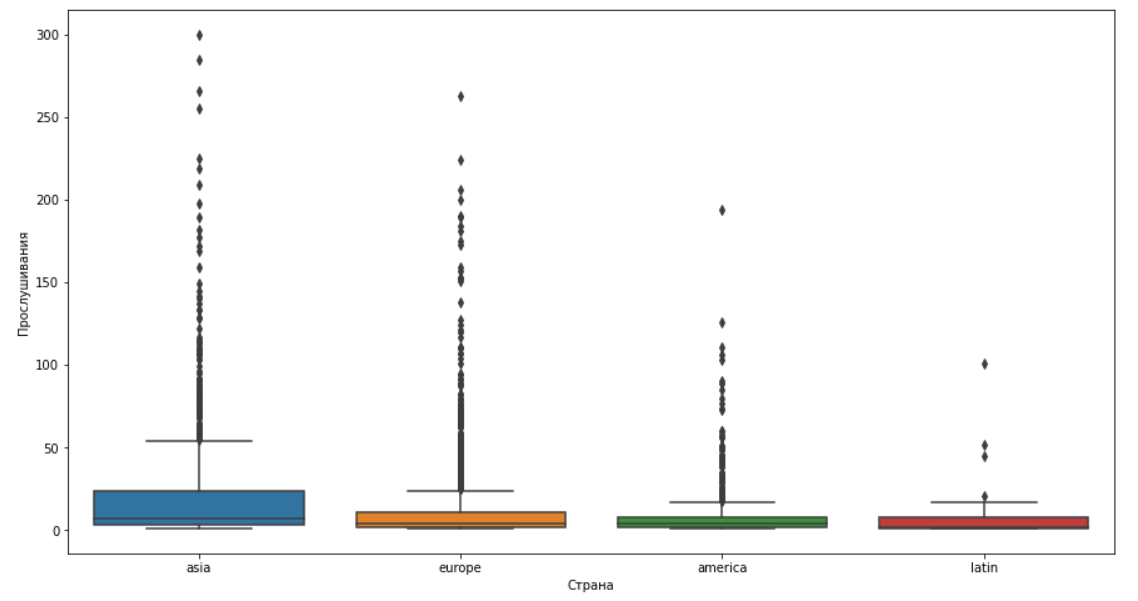


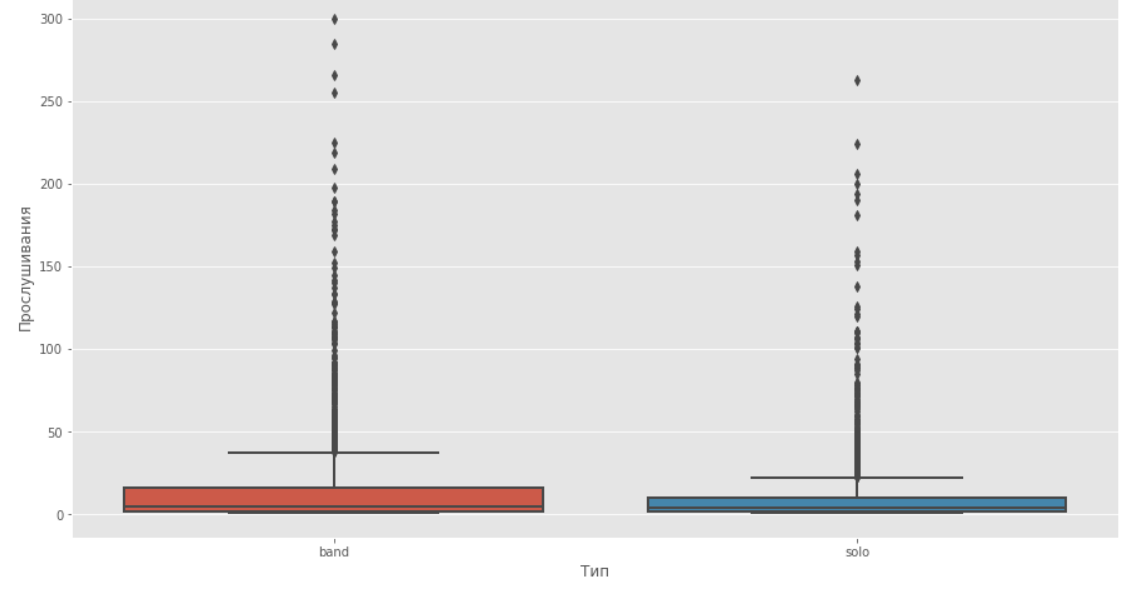
С первого взгляда кажется, что необходимо работать с выбросами. Так как данные сами по себе разрозненные и специфичны по своей идее (есть треки, количество прослушиваний которых в десятки раз больше, чем прослушивание других из-за «заедания» и подобного), было принято решение предсказывать не возможное количество скробблов, а принадлежность к классу – будет ли этот трек прослушиваться пользователем или нет.

Пока рассмотрим другие графики, распределение по жанрам, странам и типу исполнителя:



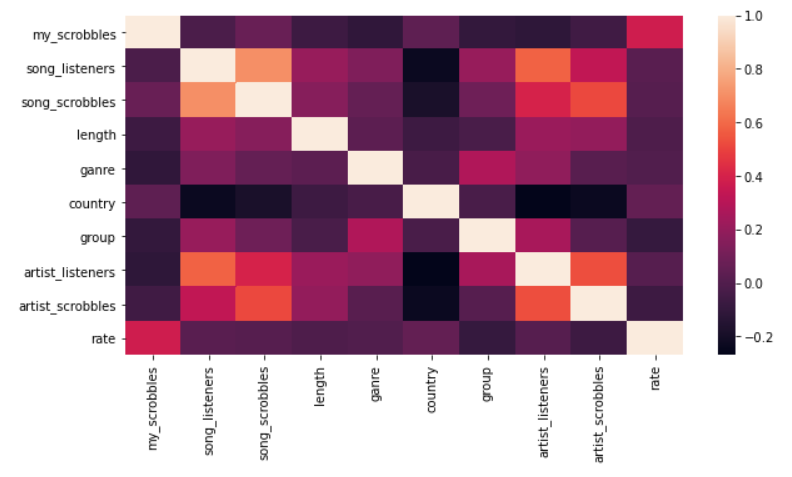
Даже по количеству выбросов заметна разница в прослушивании различных жанров.



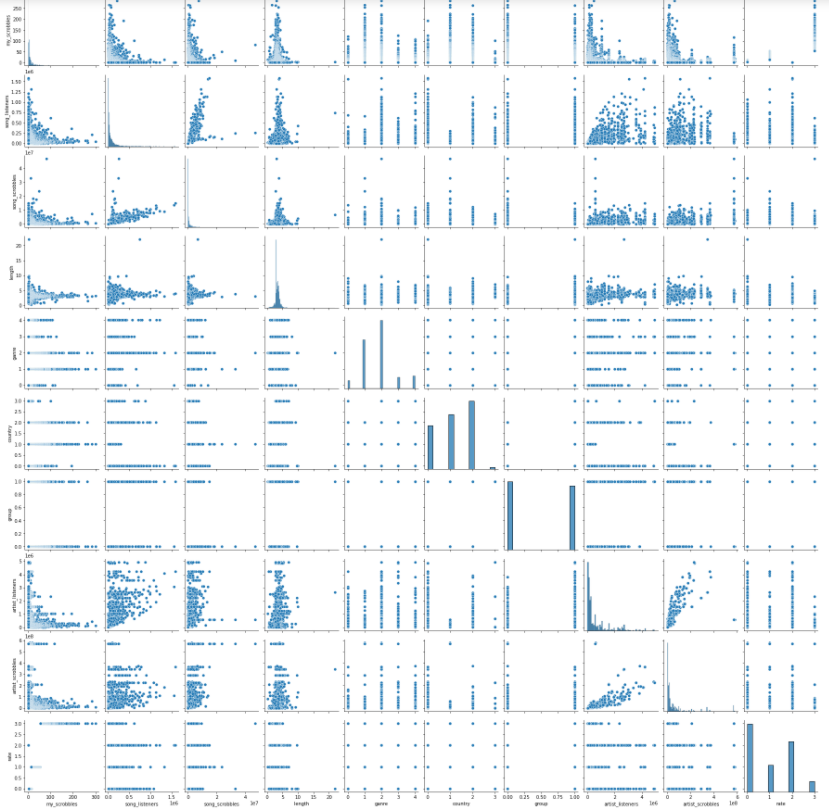


Итак, для кодирования категориальных данных в качестве первого способа был применен метод label\_encoder.

На диаграмме корреляций наблюдается стабильная зависимость между определенными факторами, что говорит о связности данных. Огромной корреляции между полями нет, что хорошо.



Также были построены графики зависимостей различных факторов относительно друг друга. В большинстве своем они не несут огромной информативной ценности, однако благодаря некотором из них можно проследить тренд зависимости.

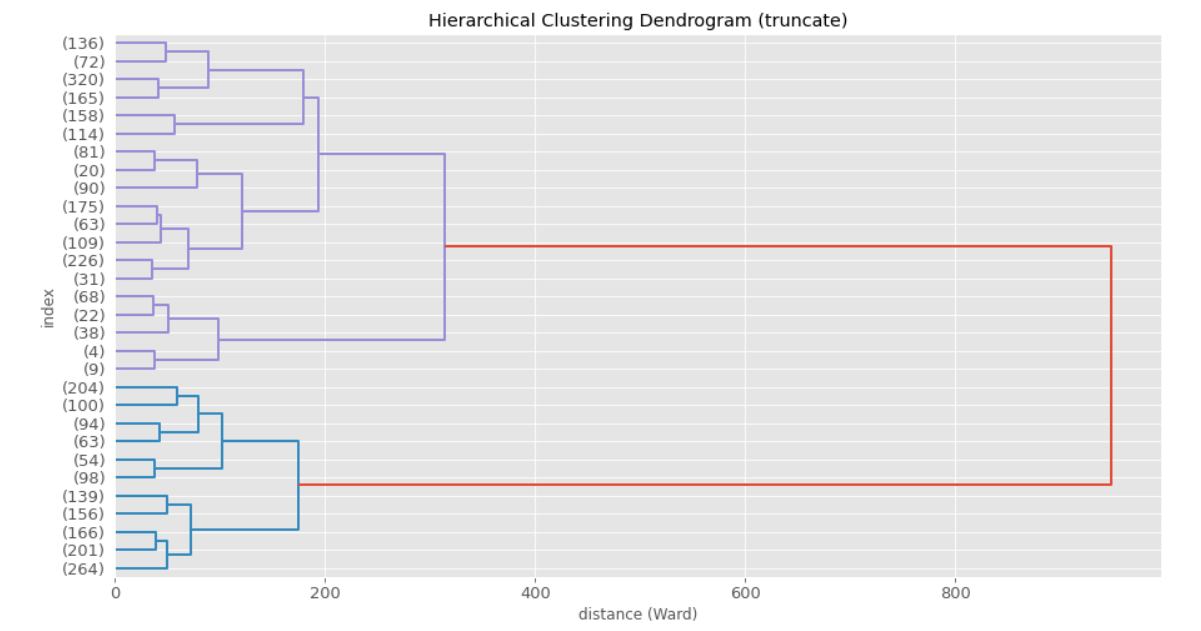


# **Кластерный анализ**

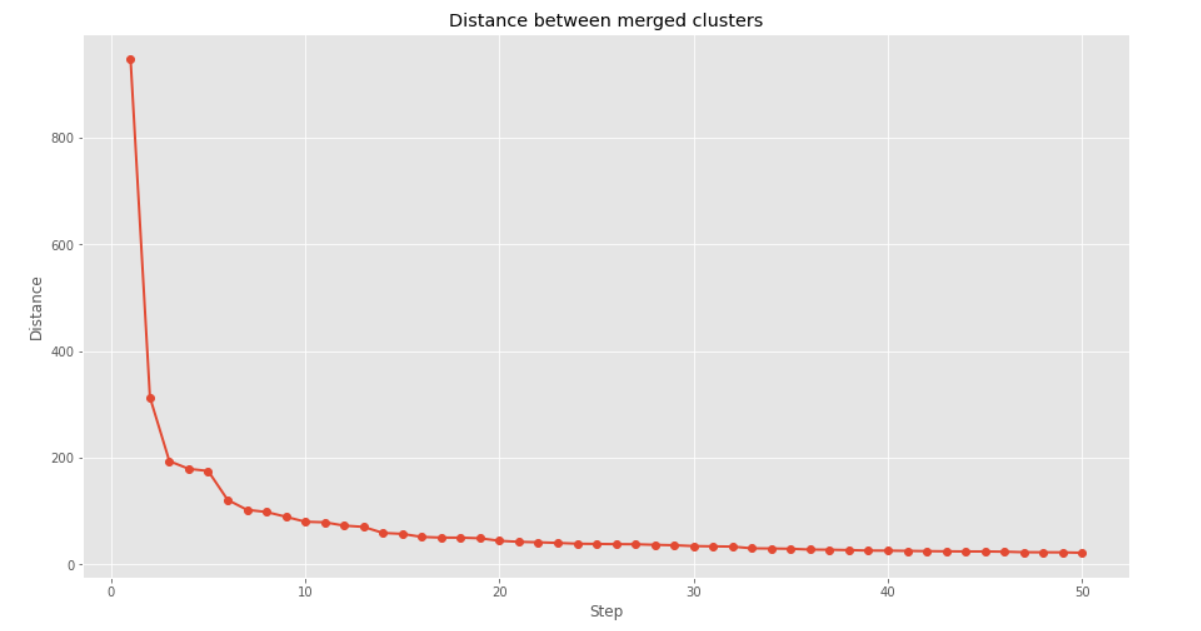
Изначально для анализа было сформировано четыре класса. Они были сформированы по количеству скробблов пользователя относительно друг друга. Распределение вышло следующим:



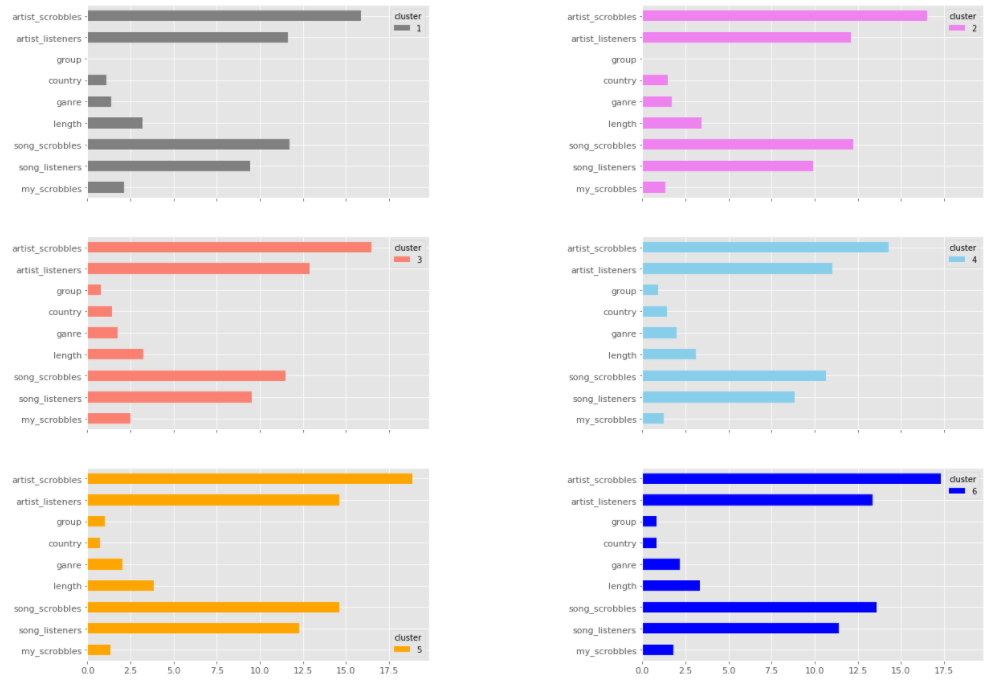
Однако нельзя полагаться лишь на визуальное представление данных, поэтому был проведен кластерный анализ. Дендограмма выглядит следующим образом:



«Каменистая осыпь»



По данным графика можно сделать вывод, что оптимальными могут быть 3-4 или 6 кластеров. При рассмотрении каждого из вариантов, получилась следующая картина:

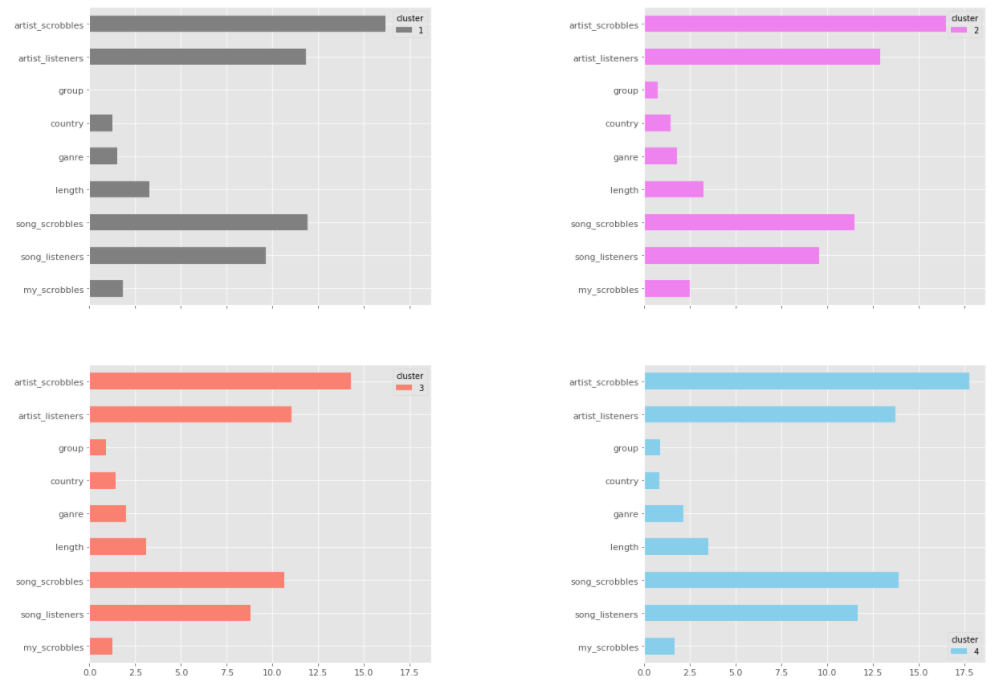


Сначала рассмотрим первый вариант – 6 кластеров.

Каждый из кластеров характеризует вид/жанр трека, а также заинтересованность пользователя в нем.

* Кластер 1 описывает что-то похожее на классические треки жанра k-pop с исполнителями-группами и небольшой длиной трека. Пользователь в них весьма заинтересован.
* Кластер 2 в целом схож с кластером один, за исключением «жанровости» - данные треки можно больше отнести к жанру pop, пользователь менее заинтересован в этих треках.
* Кластер 3 описывает отношение к трекам жанров pop и rap. Здесь должны быть треки «чартов», что имеют широкую популярность
* Кластер 4 практически полностью дублирует кластер 3 за изменением типа исполнителя – почти стопроцентный соло.
* Кластер 5 – соло-исполнители из стран америки в жанре rap.
* Кластер 6 – почти полностью дублирует кластер 5 за исключением того, что жанр больше приближен к rock.

Итак, можно сделать вывод, что 6 кластеров – слишком много, так как они начинают дублировать друг друга. Здесь будет тяжелее определить их смысловую нагрузку в разрезе заинтересованности пользователя.

Рассмотрим второй вариант – 4 кластера, как и было выделено в начале (3 кластера опустим, так как там совсем «плохое» распределение по количеству треков – в двух кластерах больше 1000, в одном меньше 80

* Кластер 1 описывает что-то похожее на классические треки жанра k-pop с исполнителями-группами и небольшой длиной трека. Пользователь в них весьма заинтересован.
* Кластер 2 – соло-исполнители и группы из различных стран в жанрах pop, пользователь очень в них заинтересован.
* Кластер 3 – соло-исполнители скорее всего в жанре rap и pop, могут относиться к европейским странам. Это не чартовая музыка для узкого круга слушателей. Пользователь чуть менее заинтересован в такой музыке.
* Кластер 4 относит данные к трекам с большим уклоном от pop к другим жанрам, а страны – скорее всего америка и европа. Пользователь в них заинтересован средне.

По получившимся данным принимаем – 4 кластера лучше, чем 6, потому что они имеют различный характер и заметно отличаются друг от друга.

# **Регрессия и классификация**

В данном разделе будут рассматриваться слежующие виды регрессий:

* Метод ближайших k-соседей (KNN)
* Логистическая регрессия
* SVM и как частный случай SVC – Метод опорных векторов

А также Random Forest и Extra Trees для решения задачи классификации.

Изначально задачей стояло предсказание скробблов, но после нескольких неудачных построений моделей (данные очень разрознены, а выбросы важны) – модели обучались с качеством не более, чем 18% - было принято решение перейти к задаче классификации. В случае предсказания по классическому варианту регрессии были использованы Линейная регрессия, Lasso, Ridge, а также LassoCV и RidgeCV. Самая высокая отметка качества – 18%.

Были рассмотрены два варианта для двух случаев: если классов 4, если классов 6, если данные кодированы label\_encoder и если данные были кодированы one-hot encoder.

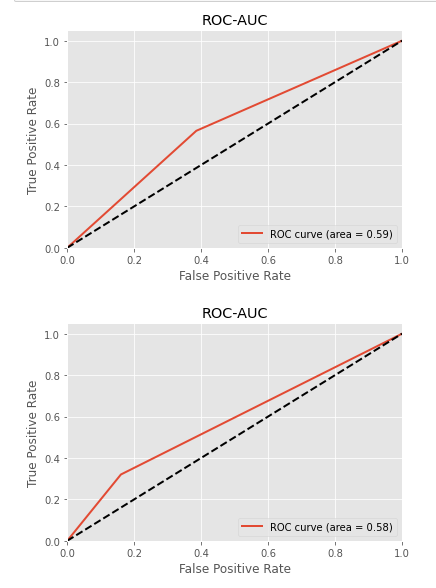
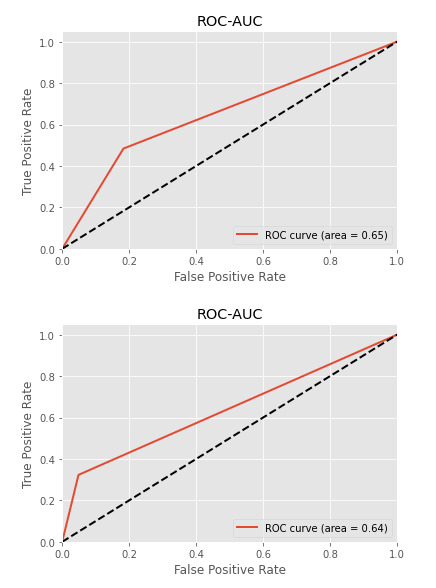
Отметим, что после принятия отрицательных результатов по 6 классам и Random Forest данные варианты не были использованы. Extra Trees использовался в самом конце после всех прогнозов и выводов, так как не рассматривался на семинарах (уверенности в применении данного метода я не чувствовала, но была привлечена его результатами)

Рассмотрим статистику по построенным моделям:

|  |  |  |  |
| --- | --- | --- | --- |
| Случай/модель | 6 классов, label | 4 класса, label | 4 класса, one-hot |
| KNN | 0,43 | 0,51 | 0,51 |
| Logistic | 0,38 | 0,47 | 0,47 |
| SVC | 0,36 | 0,47 | 0,48 |
| Random Forest | -0,05 | 0,09 |  |
| Extra Trees |  |  | 0,55 |

Кроме этого, стоит отметить, что при кодировании в one-hot в случае метода KNN целевая переменная (rate) также классифицируется, в то время как в методах Логистической регрессии, SVC и Extra Trees рассматривается как отдельное число. Также по каждому из методов был применен GridSearchCV для нахождения лучшей конфигурации модели.

Лучший из вариантов регрессий был построен методом ближайших соседей – KNN, реализуем графики ROG-AUC по его результатам:

Предсказание было построено по треку, прослушивание которого не было зарегистрировано в аккаунте пользователя. После ввода данных и прогнозирования по моделям KNN и Extra Trees результаты были следующие:

«На постоянной основе» по KNN и «Можно слушать» по Extra Trees. Результаты похожи на правду.

# **Выводы**

Данная работа претерпела множество видоизменений, начиная от постановки задачи, заканчивая методами ее реализации. Результат был получен ниже среднего качества, однако в разы лучше первоначального.

Факторы, влияющие на оценку музыку слушателем выявлены следующие: жанр, страна/язык, группа/соло, возраст исполнителя/его карьеры и прослушивание трека в мире, а также в частных случаях длина трека.

Данные, полученные парсингом сайта last.fm, были проанализированы и скорректированы согласно стандартам анализа. Для визуализации ситуации, принятия решений относительно будущих задач были представлены графики.

Благодаря кластерному анализу методом KMeans были сформированы несколько вариантов разбиения данных на категории, а также они были оценены критически.

Для построения предсказания были использованы модели KNN, SVM SVC, Логистической регрессии, а также рассмотрены варианты использования Random Forest и Extra Trees.

Качество моделей можно попробовать улучшить посредством перестроения данных с новыми классами, например, попробовать вариант с тремя классами, а также поэкспериментировать с длительностью трека и возрастом/карьерой исполнителя.