Данные для данного проекта были взяты с сайта kaggle.com

Ссылка на BibTex:

@misc{amitansh joshi_amit parolkar_vedant das_2023,
title={Spotify_1Million_Tracks},url={https://www.kaggle.com/dsv/5987852},
DOl={10.34740/KAGGLE/DSV/5987852}, publisher={Kaggle),
author={Amitansh Joshi and Amit Parolkar and Vedant Das}, year={2023}}

Данные были получены пользователем Kaggle со стриминговой музыкальной платформы Spotify, с использованием библиотеки Python Spotipy, которая позволяет пользователям получить доступ к музыкальным данным сервиса при помощи API. Набор данных, взятый изначально, включает в себя 1 миллион треков с датой выпуска от 2000 до 2023 года. Каждый из них был проанализирован по 19 переменным. Все данные были занесены в общую таблицу в формате CSV. Всего в наборе данных присутствуют треки 61 445 музыкальных исполнителей в 82 жанрах.

Изначальные цели автора набора данных: Данный набор данных, по мнению автора, предназначен для исследовательских целей. Значимость этих данных заключается в возможности выявить шаблоны пользовательского выбора и предсказать популярность песен до их непосредственной публикации на музыкальных платформах. По мнению автора, данный набор данных может быть использован для создания различных предсказательных моделей машинного и глубокого обучения с использованием различных методов и библиотек.

Мои первоначальные цели: в отличии от автора набора данных, моей главной целью являлось не создание моделей машинного обучения, а непосредственно сам анализ данных. Мне было интересно, существуют ли какие-либо взаимосвязи между переменными.

Что я для этого сделал?

Я создал файл в программе Pycharm Professional, импортировав туда три библиотеки: Pandas, главная библиотека проекта, при помощи которых и проводились все вычисления, Matplotlib и Seaborn, которые использовались для визуализации полученных данных.

Затем я решил сократить количество переменных до 7:

'artist_name' – имя исполнителя; **'track_name'** – название трека; **'year'** – год публикации песни. Данные переменные используются в качестве индикаторов, при помощи которых отслеживаются сами треки. Касаемо переменной **'artist_name'**, я решил выбрать семь случайных музыкальных исполнителей. Ими стали: Jason Mraz, Joshua Hyslop, Andrew Belle, Nicola Conte, Amon Tobin, David Gray, Harley Poe.

'popularity' – популярность песни; **'tempo'** – темп песни. Данные переменные использовались для построения первой гипотезы.

'danceability' – танцевальность трека; **'duration_ms'** – длительность трека. Данные переменные использовались для построения второй гипотезы.

!!! Переменные **'popularity'** и **'danceability'** были выбраны исключительно потому, что они являются показателями, по которым можно делить песни на менее и более успешные. Каким образом были получены эти значения, точно установить не удалось. Скорее всего, эти данные содержатся непосредственно в Spotify API.

После выбора всех переменных были сформулированы две гипотезы:

Первая гипотеза: Чем быстрее темп у трека, тем он популярнее.

Вторая гипотеза: Чем больше длительность трека, тем больше он подходит для танцев.

После формулирования гипотез при помощи конкатенации были сформированы два кластера, на основе каждого из которых обе гипотезы должны были быть подтверждены, опровергнуты, или должна была быть выявлена какая-либо иная связь между переменными, неуказанная ранее в гипотезах.

В **Первый Кластер** вошли треки музыкальных исполнителей таких, как Jason Mraz, Andrew Belle и Nicola Conte за 2008, 2011, 2014, 2017, 2018, 2020, 2021 и 2023 годы, так как именно в эти годы каждый из музыкантов выпустил хотя бы по одному треку.

Во **Второй Кластер** вошли треки музыкальных исполнителей таких, как Joshua Hyslop, Amon Tobin, David Gray и Harley Poe за 2012, 2015, 2022 годы, так как именно в эти годы каждый из музыкантов выпустил хотя бы по одному треку.

После завершения формирования двух кластеров была проведена подготовка данных для визуализации: данные в первом кластере для удобства были отсортированы по возрастанию по переменной **темп - 'tempo'**, данные во втором кластере были отсортированы по возрастанию по переменной **длительность трека - 'duration_ms'**.

Наконец, была произведена визуализация данных обоих кластеров. В **Первом Кластере** зависимой переменной была **популярность ('popularity')**, а независимой переменной – **темп ('tempo')**.

Результаты визуализации Первого Кластера:

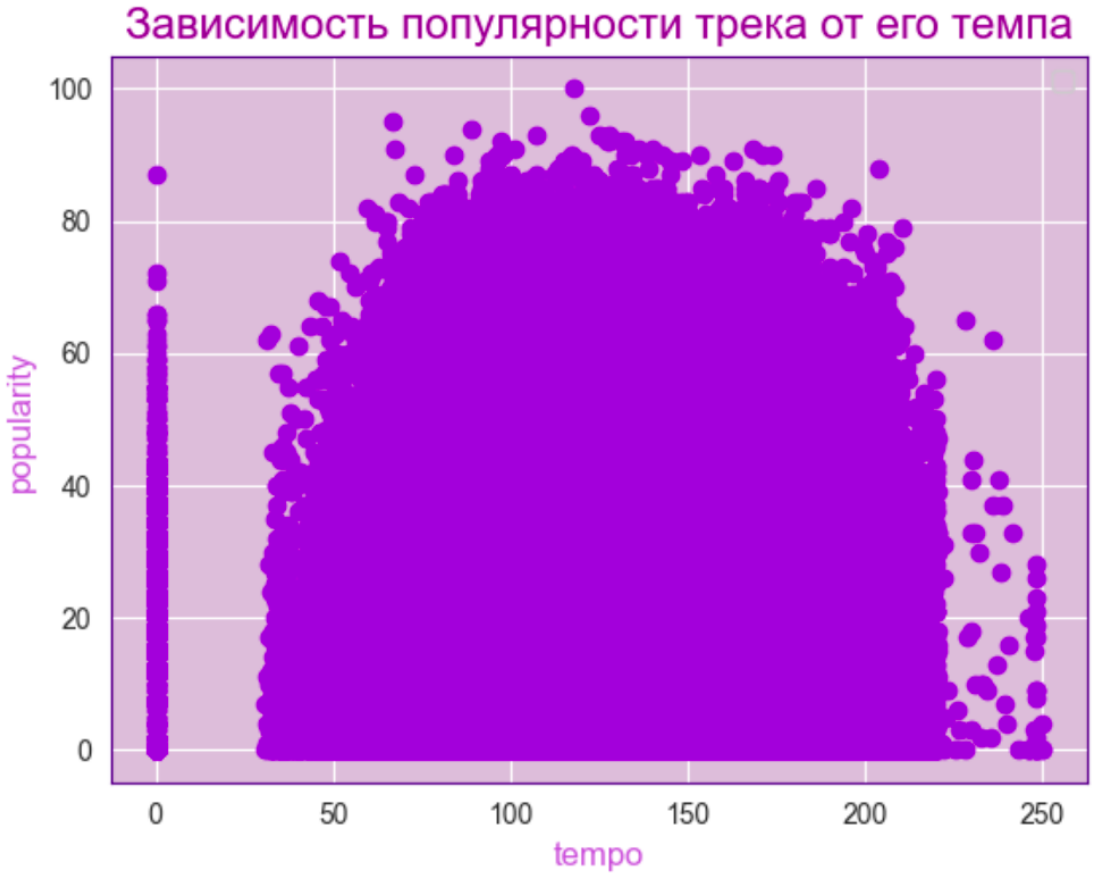
Визуализация данных из полученной выборки не позволяют сделать однозначный вывод по поводу наличия или отсутствия зависимости, так как невозможно четко проследить преобладание одного тренда над другим: точки расположены практически хаотично. Единственным интервалом, где наблюдается нечто, похожее на кластер, является область темпа в районе около 120. Поскольку нам важно проверить правдивость или ложность ранее установленной гипотезы, считаю необходимым провести визуализацию данных по тем же самым переменным, но уже по всей генеральной совокупности.

Во **Втором Кластере** зависимой переменной была **танцевальность (‘danceability’)**, а независимой переменной – **длительность трека (‘duration_ms’)**.

Результаты визуализации Второго Кластера:

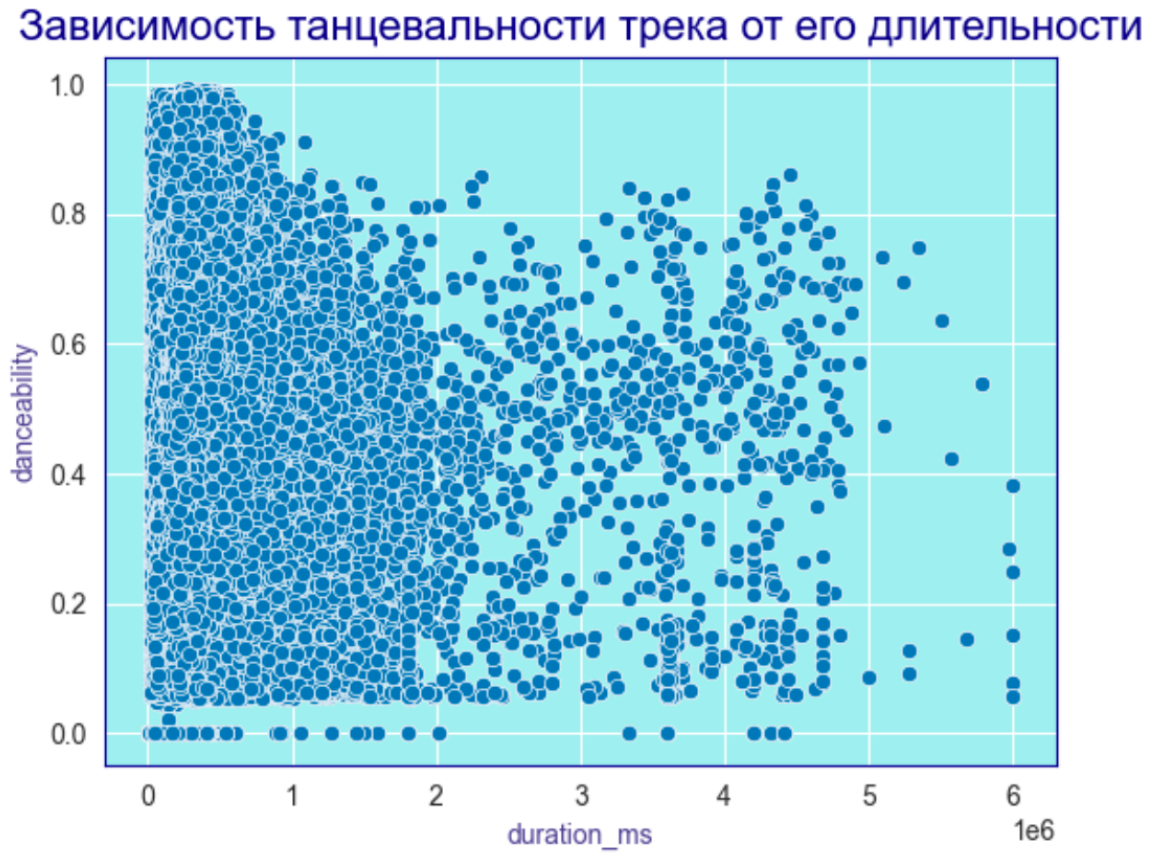
Результаты визуализации полученной выборки не позволяют сделать однозначный вывод по поводу правдивости или ложности ранее выдвинутой гипотезы. Тем не менее, по количеству пиков графика линейной зависимости можно предположить, что в интервале между 200 000 и 300 000 мс расположилось больше количество треков, которые, согласно имеющимся в выборке данным, более подходят для танцев. Тем не менее, этих данных недостаточно для однозначных выводов, поэтому считаю необходимым протестировать переменные на всей генеральной совокупности.

Результаты визуализации при проверке гипотезы для Первого Кластера:



В генеральной совокупности преобладают треки со средним темпом 125. Исходя из получившейся точечной диаграммы можно понять, что **НАИБОЛЕЕ ПОПУЛЯРНЫЕ ТРЕКИ** написаны в **СРЕДНЕМ ТЕМПЕ**, следовательно, гипотеза о том, что при повышении темпа трека растет его популярность, **НЕВЕРНА.** Однако диаграмма генеральной совокупности более информативна, чем диаграмма выборки.

Результаты визуализации при проверке гипотезы для Второго Кластера:



В результате визуализации переменных **«danceability»** и **«duration_ms»** на всей генеральной совокупности **НЕ ОБНАРУЖЕНО НИКАКОЙ ЗНАЧИМОЙ ЗАВИСИМОСТИ МЕЖДУ ДВУМЯ ПЕРЕМЕННЫМИ**. Возможно, между ними и существует какая-то связь, но гипотеза, выдвинутая ранее, является **ЛОЖНОЙ**, так как она не получила подтверждение.

ВЫВОД: Данный набор данных несмотря на то, что ни одна гипотеза не получила подтверждение, тем не менее оказался полезен, так как была выявлена взаимосвязь между темпом трека и его популярностью. Также данный набор данных можно визуализировать различными способами, такими как линейная или точечная диаграмма.