Правительство Российской Федерации
Федеральное государственное автономное
образовательное учреждение высшего
профессионального образования
Национальный Исследовательский Университет «Высшая Школа
Экономики»

Экономика

Домашнее задание № 2 по предмету «Эконометрика-1»

На тему: «Влияние плохих и хороших слов на популярность песни»

Работу выполнили:
Сафарова Рината
Гадаева Медина
Григорян Тигран
Преподаватель:
Бывальцева Станкевич Анастасия Александровна

Москва
2023

Введение

В современном мире анализ текстов песен становится все более важным, особенно в контексте эконометрических исследований. Наш проект направлен на использование методов эконометрики для выявления закономерностей взаимосвязей между текстами песен и их популярностью трех известных исполнителей: Дрейка, Эминема и Рианны. Выбор данной темы обоснован желанием проверить то, как именно слова и тот кто их озвучил влияют на популярность песни.

Цель и задачи исследования

Цель нашего исследования заключается в выявлении влияния различных параметров текстов песен на их популярность, измеряемую количеством просмотров. Для достижения этой цели мы ставим перед собой следующие задачи:

1. Собрать и структурировать данные о текстах песен, включая информацию о альбомах, песнях, текстах, просмотрах, годе выпуска и других параметрах. Наши данные были отобраны и загружены всеми участниками команды 21.11.2023 с сайта Kaggle.

2. Добавить новые параметры в датасеты, такие как количество "плохих"и "хороших"слов в песнях, количество слов в названии песни, среднее количество букв в слове, корреляция слов в песне с самой популярной песней исполнителя и другие.

3. Провести анализ влияния добавленных параметров на популярность песен с использованием методов эконометрики в Python.

4. Визуализировать данные для выявления зависимостей и паттернов.

Актуальность исследования

Научная литература предоставляет ценные исследования, которые проливают свет на факторы, влияющие на популярность музыкальных произведений. Анализ рейтингов Billboard с 2014 по 2016 годы, проведенный исследователями в семи жанрах, выявил несколько важных закономерностей. Например, необычные песни, несовпадающие со стандартными категориями, имеют больший потенциал стать популярными. Это вывод основан на анализе 4 200 песен, подразделенных на категории.

Результаты также указывают на то, что факторы, такие как радиовещание, исполнитель, количество слов и лингвистический стиль, не оказывают существенного влияния на популярность песен. Это важное открытие может помочь лучше понять, какие аспекты текстов музыки действительно привлекают внимание слушателей.

Исследователи из Университета Кертина (Австралия) изучили 271 песню, с 1999 по 2014 год появлявшуюся в британском чарте. Их интересовала беглость обработки текста песни, то есть насколько легко и быстро наш мозг обрабатывает предлагаемую информацию. Это неминуемо влияет на то, как мы эту информацию воспринимаем — «слишком сложно и занудно» или «просто и зажигательно».Тексты песен обрабатывала компьютерная программа, которая оценивала их удобочитаемость и общую сложность текста. Популярность исследователи оценивали по тому, какую позицию и как долго песня занимала в еженедельных чартах.

Оказалось, чем проще мозг слушателя воспринимал текст песни, тем большей популярности она добивалась, занимая наивысшие позиции в чартах.

Другое исследование, проведенное сравнением российской рэп-сцены с использованием методов Text Mining и R, позволяет взглянуть на вопрос более широко. Анализ текстов песен российских рэп-исполнителей, включая Noize Mc, Kasta, Pharaoh и Morgenshtern, подчеркивает различия в языке, словарном запасе и эмоциональной окраске между артистами. Стремление к инновациям, подкрепленное желанием сохранить знакомость и приятность, становится ключевым фактором успешности песен.

Таким образом, наше исследование нацелено на применение улучшение понимания влияния различных параметров текстов на популярность музыкальных произведений.

Источники данных

- **Эминем:** Данные о текстах песен Эминема были собраны с сайта Kaggle. Источник данных: `https://www.kaggle.com/datasets/aditya2803/eminem-lyrics/code`
- **Рианна:** Информация о текстах песен Рианны была собрана и структурирована участниками команды вместе с данными, загруженными с Kaggle. Источник данных: `https://www.kaggle.com/datasets/vivovinco/rihanna-lyrics/data`
- **Плохие слова:** Данные о "плохих"словах были взяты из набора данных на Kaggle. Источник данных: `https://www.kaggle.com/datasets/nicapotato/bad-bad-words`
- **Дрейк:** Информация о текстах песен Дрейка была собрана с Kaggle. Источник данных: `https://www.kaggle.com/datasets/juicobowley/drake-lyrics`
- **Позитивные слова:** Набор положительных слов был взят из репозитория на GitHub. Источник данных: `https://gist.github.com/mkulakowski2/4289437`

Экономическая модель

В центре нашего исследования стоит модель, предназначенная для анализа взаимосвязей между текстовыми параметрами песен и их популярностью, измеряемой в просмотрах. Для построения модели линейной регрессии оцененной методом наименьших квадратов, мы использовали различные методы эконометрики, проводя ряд анализов и тестов, которые мы рассмотрим чуть позже.

Список объясняющих переменных:

1. **Количество "плохих"слов в песне (bad_word_count)**
2. **Количество "хороших"слов в песне (positive_word_count)**
3. **Соотношение уникальных слов в тексте песни (unique_words_ratio) (в долях)**
4. **Длина текста песни (song_length)**
5. **Длина заголовка песни (song_title_length)**

6. **Средняя длина слов в тексте песни (average_word_length)**
7. **Корреляция слов в песне с самой популярной песней исполнителя (corr_with_banger)**
8. **Пол исполнителя (sex) (0-женщина; 1-мужчина)**
9. **Раса исполнителя (race) (0-темнокожий; 1-светлокожий)**
10. **Возраст исполнителя на момент выпуска трека (age_at_release) (в годах)**

Итого в нашем датасете 532 наблюдения, 10 объясняющих переменных и 1 константная

Почему был выбран именно этот набор объясняющих переменных?

После проведенного анализа научной литературы, у нашей команды возникло предположение о том, что текстовые параметры песен, такие как выбор слов, их разнообразие, длина и структура, могут влиять на их популярность. Пол, раса и возраст исполнителя также могут оказывать влияние на восприятие и успех песен. Например:

- **Количество "плохих"и "хороших"слов:** Предполагается, что использование эмоционально окрашенных слов может привлечь внимание слушателей.
- **Уникальность текста:** Нами было замечено, что, менее уникальные тексты, с часто повторяющимися словами в наши дни имеют довольно высокую популярность.
- **Длина текста и заголовка:** На сегодняшний день, песни с несодержательным текстом имеют большую популярность, чем глубокосмысловые.
- **Средняя длина слов:** Часто людьми легче воспринимаются тексты с меньшим количеством слов.
- **Корреляция с самой популярной песней:** Вероятно, что если новая песня поддается схожей структуре с успешной предыдущей, это может повлиять на ее популярность.
- **Пол, раса и возраст исполнителя:** Личные характеристики исполнителя могут создавать определенное восприятие и влиять на предпочтения аудитории.

Содержательные гипотезы:

1. **Гипотеза 1:** Для исполнителей представляющих разные социальные/этнические группы существуют свои стереотипы и образы, принадлежность которым может влиять на факторы успеха. проверим гипотезу о том, что влияние взятых нами факторов на популярность песни для мужских и женских исполнителей - используем для этого тест Чоу, взяв за уровень значимости 5%.
2. **Гипотеза 2:** Мы предпологаем, влияние положительных и отрицательных словах по модулю одинаковое. Для этого мы реализуем функцию, которая выполнит T-тест для значений по модулю и без модуля.

3. **Гипотеза 3:** Мы предполагаем, что чем выше средняя длина слова в тексте песни, тем она менее популярна. Данное предположение возникло после анализа исследования проведенного университетом Кертина, ведь они, как раз, подчеркивают значимость "формата"слов, его простоту и другие факторы.
4. **Гипотеза 4:** Мы предполагаем, что распределение просмотров по исполнителям одинаковое

Предварительный анализ данных

Распределение признаков регрессоров: Рис. 1

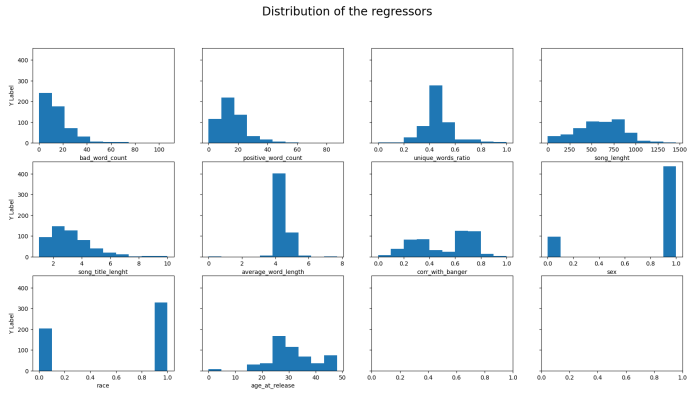


Рис. 1: Рис. 1

Влияние категориальных признаков: Рис. 2

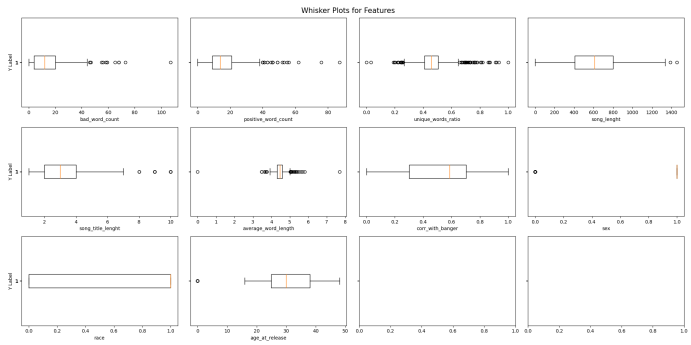


Рис. 2: Рис. 2

В общем, все регрессоры имеют достаточно «гладкое» распределение без выбросов. Ключевые переменные (хорошие и плохие слова) имеют распределение похоже на экспоненциальное.

Это означает, что в большинстве случаев значения этих переменных достаточно низкие, то есть количество "негативных"и "позитивных"слов в песнях обычно невелико. Также стоит отметить, что исходя из графика средняя длина слова в тексте песен составляет 4 буквы, что говорит о том, что большинство песен исполнителей достаточно просты для восприятия.

Box plots для категориальных переменных: Рис. 3

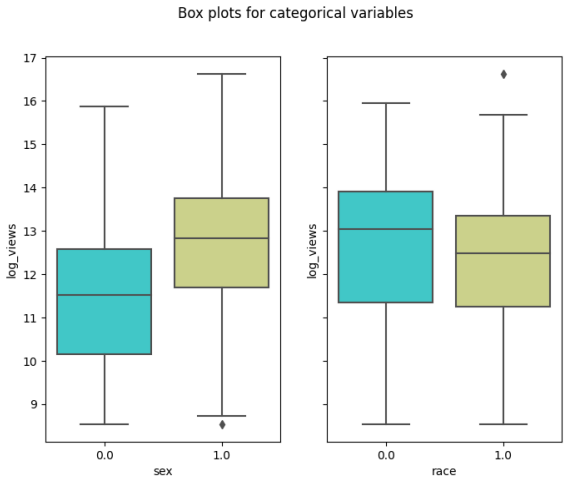
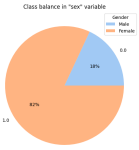


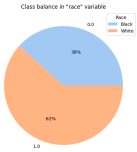
Рис. 3

Мы можем наблюдать, что пол исполнителя оказывает влияние на распределение целевой переменной (мужчины-исполнители имеют более высокое число просмотров), в то время как цвет кожи оказывает значительно меньшее влияние. Давайте посмотрим на баланс классов в категориальных переменных

Баланс классов в категориальных переменных: Рис. 4 и 5



(a) Рис. 4



(b) Рис. 5

В целом, баланс классов сохраняется

Оценка модели регрессии для предсказания популярности песен

Для анализа взаимосвязи между текстовыми параметрами песен и их популярностью, измеряемой в просмотрах, была построена регрессионная модель. Мы провели анализ трех различных вариантов моделирования с использованием различных методов масштабирования. Перед тем как начать наш анализ, мы внесли корректировки в данные, поскольку создатель датасета подтвердил, что к некоторым песням отсутствовал доступ для извлечения информации, что привело к частичному отсутствию данных. В связи с этим информация о треках, для которых не удалось получить данные, была заполнена нулевыми значениями в колонке "views". Исходя из этого, у нас есть законное основание исключить эти треки из нашего исследования. Давайте рассмотрим результаты нашей модели линейной регрессии оцененной методом наименьших квадратов.

(a) note scaled

	Dep. Variable: log_views					
	coef	std err	t	P>|t|	[0.025	0.975]
const	8.2759	0.780	10.615	0.000	6.744	9.807
bad_word_count	0.0024	0.006	0.398	0.691	-0.009	0.014
positive_word_count	0.0011	0.006	0.168	0.867	-0.011	0.013
unique_words_ratio	-1.8139	0.801	-2.264	0.024	-3.388	-0.240
song_length	0.0017	0.000	3.870	0.000	0.001	0.003
song_title_length	-0.1075	0.039	-2.751	0.006	-0.184	-0.031
average_word_length	0.5717	0.179	3.198	0.001	0.220	0.923
corr_with_banger	1.0966	0.597	1.838	0.067	-0.076	2.269
sex	2.0364	0.222	9.183	0.000	1.601	2.472
race	-2.4639	0.240	-10.277	0.000	-2.935	-1.993
age_at_release	0.0319	0.008	4.015	0.000	0.016	0.047

(b) Table 1

(c) minmax scaled

	Dep. Variable: log_views					
	coef	std err	t	P>|t|	[0.025	0.975]
const	8.1684	0.772	10.587	0.000	6.653	9.684
bad_word_count	0.2546	0.640	0.398	0.691	-1.002	1.511
positive_word_count	0.0921	0.549	0.168	0.867	-0.987	1.172
unique_words_ratio	-1.8139	0.801	-2.264	0.024	-3.388	-0.240
song_length	2.5061	0.648	3.870	0.000	1.234	3.778
song_title_length	-0.9674	0.352	-2.751	0.006	-1.658	-0.276
average_word_length	4.3949	1.374	3.198	0.001	1.695	7.095
corr_with_banger	1.0966	0.597	1.838	0.067	-0.076	2.269
sex	-2.4639	0.240	-10.277	0.000	-2.935	-1.993
race	2.0364	0.222	9.183	0.000	1.601	2.472
age_at_release	1.5304	0.381	4.015	0.000	0.782	2.279

(d) Table 2

Таблица 1: standart scaled

	Dep. Variable: log_views					
	coef	std err	t	P>|t|	[0.025	0.975]
const	12.4466	0.057	219.294	0.000	12.335	12.558
bad_word_count	0.0309	0.078	0.398	0.691	-0.122	0.183
positive_word_count	0.0115	0.068	0.168	0.867	-0.123	0.146
unique_words_ratio	-0.2097	0.093	-2.264	0.024	-0.392	-0.028
song_length	0.4573	0.118	3.870	0.000	0.225	0.689
song_title_length	-0.1820	0.066	-2.751	0.006	-0.312	-0.052
average_word_length	0.2103	0.066	3.198	0.001	0.081	0.340
corr_with_banger	0.2381	0.130	1.838	0.067	-0.016	0.493
sex	-1.1980	0.117	-10.277	0.000	-1.427	-0.969
race	0.7831	0.085	9.183	0.000	0.616	0.951
age_at_release	0.2870	0.071	4.015	0.000	0.147	0.427

Мы с командой провели Min-Max Scaling и Standard Scaling и заметили, что результаты не сильно отличаются, поскольку обе техники масштабирования используются для установления одинакового диапазона значений. Переменные `unique_words_ratio`, `song_lenght`, `average_word_length`, `sex`, `race`, и `age_at_release` имеют статистически значимый вклад в модель. А также дабы убедиться в том что мультиколлениарности нет, мы проанализировали хэшмап таблицу для выявления значений корреляции между параметрами модели (Рис. 6).

Далее выполнили тест Рамсея, дабы убедиться, что нам не стоит в модель параметры в нейлинейном виде. Тест RESET провели с использованием функции `reset_ramsey` на модели, построенной на исходных данных с степенью свободы 5.

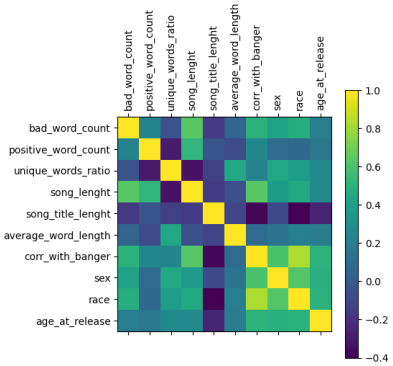


Рис. 6: Corr matrix

Таблица 2: Test Predictions and Differences

Test	Predicted	Observed	Difference
313	12.518691	12.931924	-0.413232
457	11.378251	10.729365	0.648885
476	10.392068	11.185930	-0.793862
2	11.773750	14.184884	-2.411134
108	10.098232	10.841335	-0.743103

Результаты теста Рамсея: Таблица 3

Таблица 3: Ramsey Test Results

Test Statistic	F Value	p-value
0.5366	0.7089	

Результаты Колмогорова-Смирнова: Таблица 5

Тест Колмагорова-Смирнова дал понять по p-value,что гипотеза о схожести распределения просмотров для исполнителей отвергается для всех пар исполнителей

Затем пользуясь методом Монте-Карло, мы построили модель 1000 раз на 250 случайно взятых песнях, проверили R-квадрат и вывели границы 2.5% и 97.5% квантилей из этого массива.

По сути, это позволяет оценить стабильность и вариабельность R-квадрата в зависимости от изменения данных. Если доверительный интервал узкий, это может говорить о стабильности модели на различных подвыборках данных. Провели гипотезы о равенстве параметров модели друг другу с использованием t-test и визуализировали полученный результат.

Тест	Результаты
Тест 1	t-stat = 0.4814, p-value = 6.67×10^{-18}
Тест 2	t-stat = 0.7030, p-value = 6.90×10^{-27}
Тест 3	t-stat = 0.3016, p-value = 2.31×10^{-7}

Таблица 4: Результаты теста Колмогорова-Смирнова

Проверим гипотезу о том, что влияние взятых нами факторов на популярность песни для мужских и женских исполнителей: Рис. 7, табл 6 и 7 см. приложение

Для этого мы использовали тест Чоу и взяли уровень значимости равный 5% и протестировали гипотезу для мужчин и женщин. Итоговый результат вы можете увидеть на Рис.7 (см. приложение). Как можно заметить, значение полученной f-статистики (9.77) сильно выше f-критического - гипотеза об однородности взаимосвязей для мужчин и женщин отвергается. Вероятно, такой результат обусловлен тем, что на популярность песен мужских и женских исполнителей могут влиять разные факторы в разной степени - например, для женщин число позитивных слов или возраст может играть большую роль чем для мужчин.

После полученных резудьтатов мы оценили модели по двум выборкам две модели регрессии для мужчин и для женщин (Табл 8 и 9 см. приложение). Сравнение регрессий полученных для мужчин и женщин отдельно показывает что есть разница во влиянии регрессоров - возраст для женщин оказывает большее влияние (значение коэффициента 0.4 против 0.2 у мужчин) на популярность песни, при этом длина названия песни и средняя длина слова в текстах у женщин оказалось незначимым коэффициентом в отличие от мужчин, что возможно обусловлено спецификой выборки (часть датасета составляют песни рэп-исполнителя Эминема);

Теперь рассмотрим наше предположение о том, что чем выше средняя длина слова, тем популярнее песня: Табл. 1 OLS Regression Results

Коэффициент при переменной average_word_length имеет положительное значение (0.2103), следовательно, наша гипотеза о негативном влиянии длинных слов в песнях на популярность отвергается - "сложные"слова напротив, повышают популярность песни. можно предположить, что это связано с тем что слушатели предпочитают песни с более содержательной смысловой составляющей, так как такие песни вызывают больший эмоциональный отклик.

Проверим на равенство параметры при положительных и отрицательных словах по модулю: Рис.8 и 9 (см. приложение)

t-stat	t-crit	df
-0.00012	1.64778	522

Для большей уверенности, мы провели также T-тест, который показал в свою очередь, что нулевая гипотеза отвергается, а значит, что подозрения о незначимости модели пропадают окончательно. Мы провели T-тест отдельно для мужчин и женщин, и визуализировали полученный результат (Рис 9 и 10 см. в приложении. Как можно увидеть, наша гипотеза не отвергается потому что она попадает в промежуток между критическими значениями

Общий Вывод:

Результаты исследования показывают, что популярность песен может быть частично объяснена текстовыми параметрами и личными характеристиками исполнителя при этом характер этой взаимосвязи различается для мужчин- и женщин- исполнителей. В модели для женщин выяснилось, что число плохих слов, позитивных слов, число уникальных слов являются незначимыми параметрами, что показывает что содержание текстов песен не оказывает значительного влияния на популярность песен, при этом, как упоминалось выше, на популярность оказывает значительное влияние возраст исполнительницы. В модели для мужчин аналогично текстовые характеристики (число уникальных слов, позитивных и тд) не оказались значимыми. Также для мужских исполнителей подтвердилось предположение о том, что песни с повторяющимися словами лучше запоминаются (значение коэф. при unique words ratio отрицательно) и чем длинее песня, тем меньше ее популярность. Из этого мы можем сделать вывод, что эмоциональный окрас песни (негативный/позитивный) сам по себе не оказывает значительного влияния на популярность песни, что может быть обусловлено тем что слушатели могут предпочитать оба типа песен, в зависимости от настроения.

Используемые источники

1. *The Steppe:* https://the-steppe.com/gorod/chto-vliyaet-na-populyarnost-pesen
2. *Habr:* https://habr.com/ru/articles/501162/
3. *Towards Data Science:* https://towardsdatascience.com/do-hit-songs-have-anything-in-common-3759994
4. *GeeksforGeeks:* https://www.geeksforgeeks.org/how-to-perform-an-f-test-in-python/
5. *YouTube Video:* https://youtu.be/k9PPuTmTLMk?si=z2zPxXu1w3fq-OME&t=270
6. *Bolshoy Vopros:* http://www.bolshoyvopros.ru/questions/858638-pochemu-pesni-nesoderzhatelnym-tekst
   html

Dep. Variable:	**log_views**	**R-squared:**	0.369			
Model:	**OLS**	**Adj. R-squared:**	0.356			
Method:	**Least Squares**	**F-statistic:**	27.67			
Date:	Sun, 10 Dec 2023	**Prob (F-statistic):**	1.02e-37			
Time:	17:52:22	**Log-Likelihood:**	-699.16			
No. Observations:	436	**AIC:**	1418.			
Df Residuals:	426	**BIC:**	1459.			
Df Model:	9					
Covariance Type:	**nonrobust**					

coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**	
const	10.7224	0.831	12.900	0.000	9.089	12.356
bad_word_count	-0.0051	0.006	-0.885	0.377	-0.016	0.006
positive_word_count	-0.0129	0.008	-1.667	0.096	-0.028	0.002
unique_words_ratio	-2.1021	0.852	-2.467	0.014	-3.777	-0.427
song_length	0.0026	0.000	5.491	0.000	0.002	0.003
song_title_length	-0.1170	0.041	-2.835	0.005	-0.198	-0.036
average_word_length	0.6042	0.189	3.201	0.001	0.233	0.975
corr_with_banger	0.5348	0.625	0.856	0.393	-0.694	1.763
race	-2.3087	0.236	-9.792	0.000	-2.772	-1.845
age_at_release	0.0203	0.008	2.687	0.007	0.005	0.035

Таблица 5: OLS Regression Results for male (в таблице строка с названиями коэффициентов сдвинулась)

Dep. Variable:	**log_views**	**R-squared:**	0.592			
Model:	**OLS**	**Adj. R-squared:**	0.554			
Method:	**Least Squares**	**F-statistic:**	15.75			
Date:	Sun, 10 Dec 2023	**Prob (F-statistic):**	4.14e-14			
Time:	19:24:52	**Log-Likelihood:**	-141.59			
No. Observations:	96	**AIC:**	301.2			
Df Residuals:	87	**BIC:**	324.3			
Df Model:	8					
Covariance Type:	**nonrobust**					

coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**	
const	-1.7387	1.965	-0.885	0.379	-5.644	2.167
bad_word_count	-0.0091	0.029	-0.313	0.755	-0.067	0.049
positive_word_count	0.0107	0.009	1.208	0.230	-0.007	0.028
unique_words_ratio	0.6465	1.589	0.407	0.685	-2.513	3.806
song_length	0.0032	0.001	2.468	0.016	0.001	0.006
song_title_length	0.0045	0.074	0.062	0.951	-0.142	0.151
average_word_length	0.5481	0.339	1.616	0.110	-0.126	1.222
corr_with_banger	0.6630	1.213	0.547	0.586	-1.748	3.074
race	0	0	nan	nan	0	0
age_at_release	0.4059	0.041	9.989	0.000	0.325	0.487

Таблица 6: OLS Regression Results for female (в таблице строка с названиями коэффициентов сдвинулась)

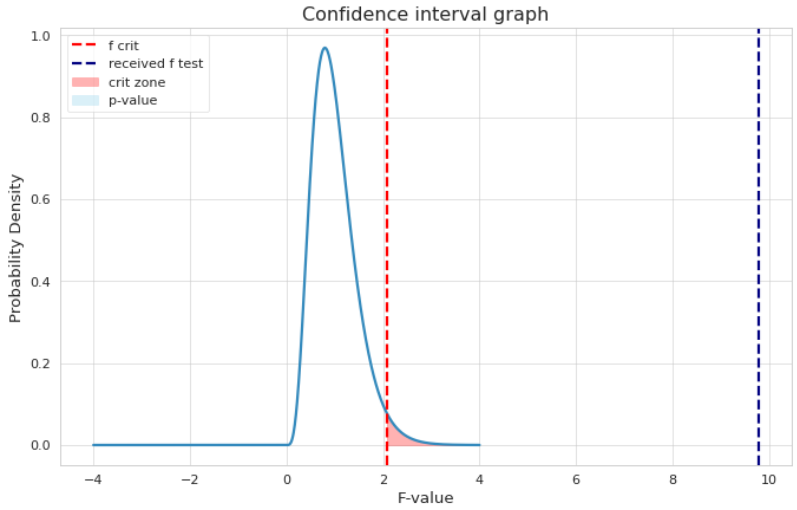


Рис. 7: график F-теста

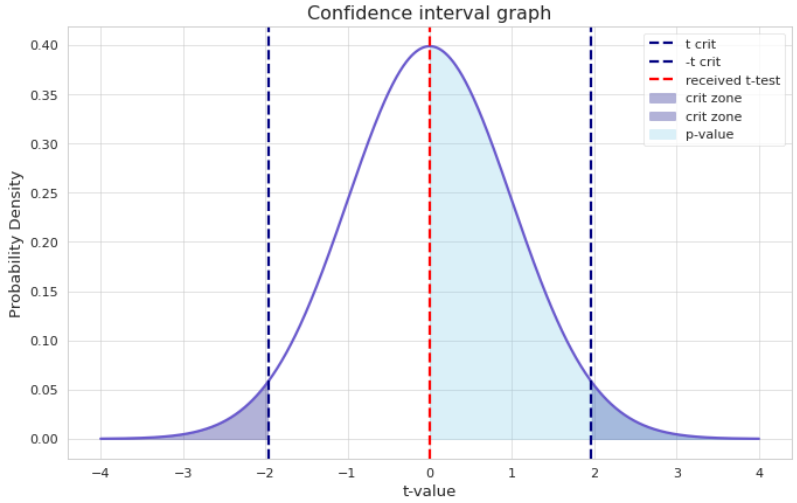


Рис. 8: График T-теста для мужчин

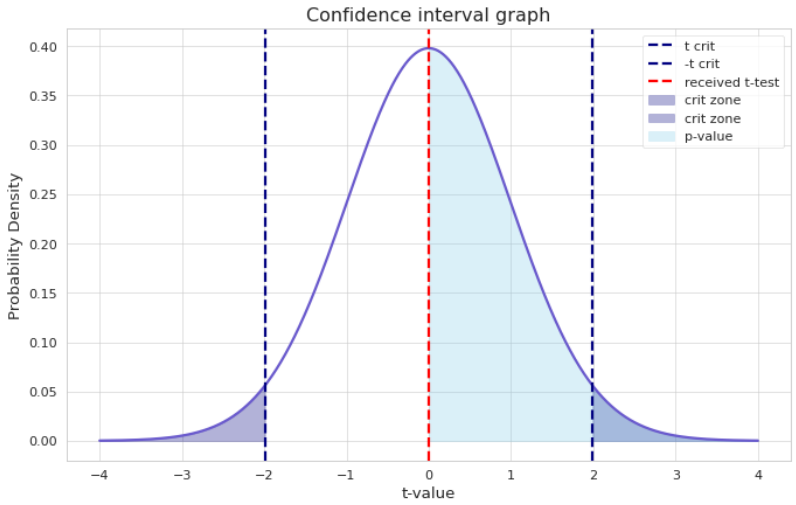


Рис. 9: График T-теста для женщин