**Звіт до лабораторної 1: Розвідковий аналіз даних**

Перелік виконавців: Руслан Дюбакін, Пріхно Ілля, Петриченко Нікіта, Резниченко Єлизавета.

**Вступ**

У цьому звіті ми представляємо результати аналізу набору даних треків Spotify у 125 різних жанрах. Наше дослідження було мотивоване бажанням відповісти на деякі значущі та цікаві дослідницькі питання, які могли б зацікавити читачів. Зокрема, ми намагалися з'ясувати зв'язок між різними аудіо характеристиками треків та їхньою популярністю, а також дослідити інші фактори, які можуть впливати на загальний досвід прослуховування користувачами. Нижче приведений повний список питань:

* Залежність популярності треку за валентністю(valence)
* Як тривалість треку впливає на його енергійність
* Як ненормативна лексика й темп впливає на мовність(speechiness) треку
* Який альбом є найпопулярнішим, що є спільного між його треками

Щоб спрямувати наш аналіз, ми сформулювали кілька гіпотез та очікувань. Наприклад, ми припустили, що валентність (позитивний чи негативний емоційний зміст) треку позитивно корелює з його популярністю. Ми також очікували, що довжина треку може впливати на його енергетику, а певна ненормативна лексика та темп можуть впливати на мовність треку.

Наше дослідження важливе, оскільки воно проливає світло на основні фактори, які впливають на популярність і загальний досвід прослуховування користувачів Spotify. Наші висновки можуть бути використані для створення рекомендаційних систем на основі вподобань користувачів або для класифікації треків на основі їхніх звукових особливостей та жанрів. Ми сподіваємося, що наше дослідження надихне на подальше вивчення та обговорення цієї цікавої теми.

У наступних розділах ми опишемо методологію, яку ми використовували для аналізу набору даних, представимо наші висновки та обговоримо їхні наслідки.

**Опис та підготовка даних**

Набір даних, використаний у нашому аналізі, було отримано з веб-сайту Kaggle.com. Це набір даних треків Spotify у 125 різних жанрах, і кожен трек має пов'язані з ним аудіо-характеристики. Дані у форматі CSV були завантажені в R за допомогою функції "read\_csv".

Перед проведенням аналізу ми виконали попереднє очищення даних. Спочатку ми завантажили набір даних і роздрукували основну інформацію про його структуру. Потім ми перевизначили типи даних деяких змінних, таких як "explicit", "track\_genre" та "mode". Ми також перевірили наявність пропущених значень і відобразили рядки з пропущеними значеннями. Нарешті, ми видалили рядки з відсутніми значеннями і повернули очищені дані.

Загалом, наш процес очищення даних був спрямований на забезпечення точності та узгодженості набору даних, що має вирішальне значення для проведення надійного аналізу. Видаливши пропущені значення та перевизначивши типи даних, ми змогли створити чистий та узгоджений набір даних, який можна використовувати для подальшого аналізу. В результаті було видалено тільки 1 запис.

Виходячи з результатів роботи функції clean\_data(), ми бачимо, що очищені дані містять 113 999 спостережень і 20 змінних. Змінні складаються з 4 символьних змінних (track\_id, artists, album\_name, track\_name), 14 числових змінних (popularity, duration\_ms, danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo та time\_signature), 2 категоріальні типи (track\_genre, mode) та 1 логічної змінної (explicit).

Зведена статистика очищених даних показує мінімальне, максимальне, медіанне, середнє та квартильне значення для кожної змінної.

Змінна популярності коливається від 0 до 100, із середнім значенням 33,24 та медіаною 35. Розподіл виглядає дещо зміщеним вліво, з більшою кількістю треків з низькою популярністю.

Змінна duration\_ms коливається від 8,586 до 5,237,295 мілісекунд, із середнім значенням 228,031 мілісекунд і медіаною 212,906 мілісекунд. Розподіл виглядає дещо зміщеним вправо, з більшою кількістю довших треків.

Змінна танцювальності коливається від 0 до 0,985, із середнім значенням 0,5668 та медіаною 0,58. Розподіл виглядає дещо зміщеним вліво, з більшою кількістю менш танцювальних треків.

Змінна енергійності коливається від 0 до 1, із середнім значенням 0,6414 та медіаною 0,685. Розподіл виглядає дещо зміщеним вліво, з більшою кількістю менш енергійних треків.

Змінна гучності коливається від -49,531 до 4,532, із середнім значенням -8,259 і медіаною -7,004. Розподіл виглядає дещо зміщеним вправо, з більшою кількістю треків з меншою гучністю.

Змінна темпу коливається від 0 до 243,37, із середнім значенням 122,15 і медіаною 122,02. Розподіл виглядає приблизно симетричним.

Змінна time\_signature коливається від 0 до 5 із середнім значенням 3,904 та медіаною 4. Розподіл виглядає дещо зміщеним вліво, оскільки більша кількість треків має часову сигнатуру 4.

Змінна track\_genre має 7 унікальних значень, причому більшість треків (107 999) належать до категорії "Інше". Решта жанрів мають по 1,000 треків.

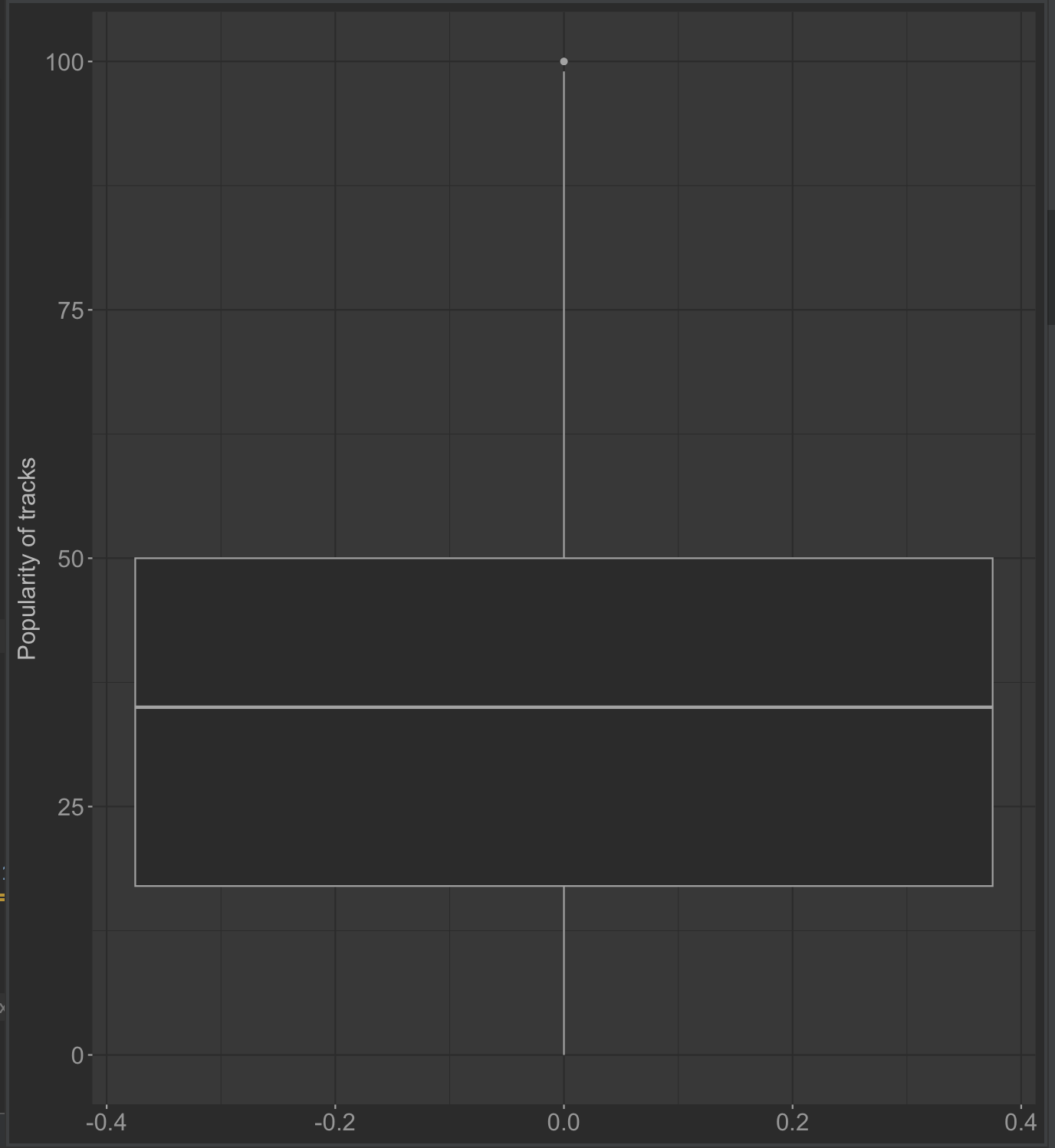
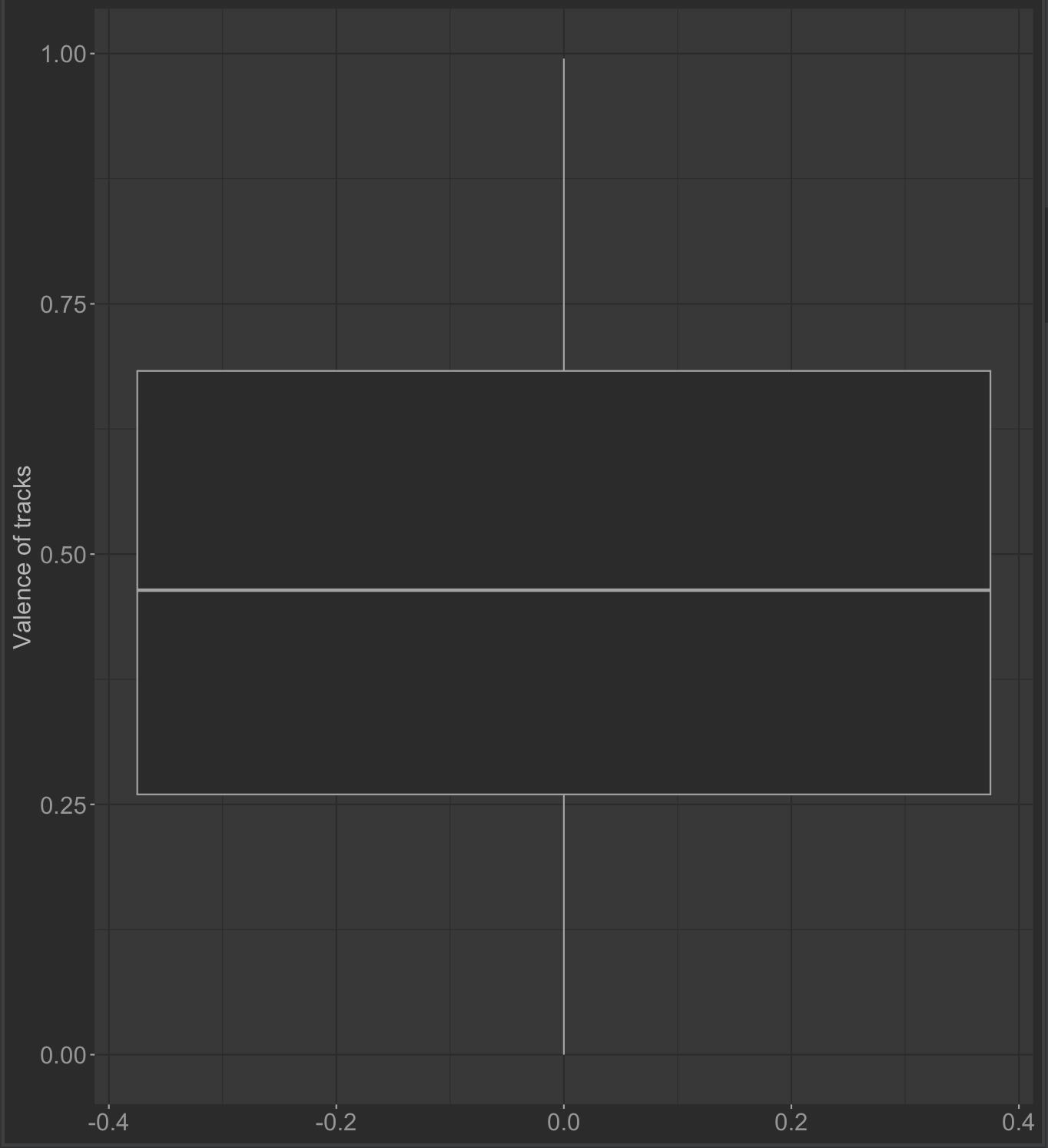
**Результати EDA у вигляді графіків**

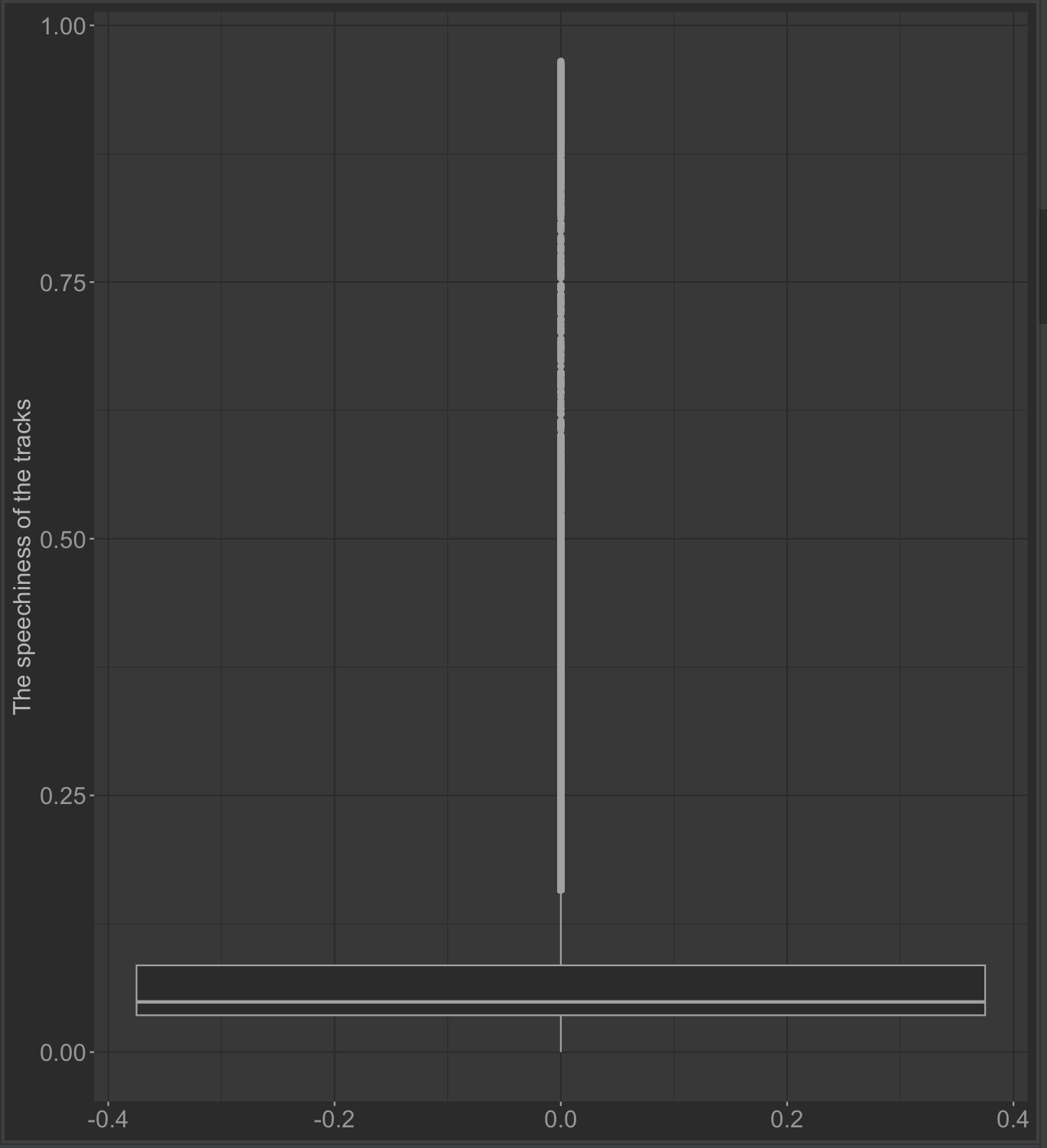
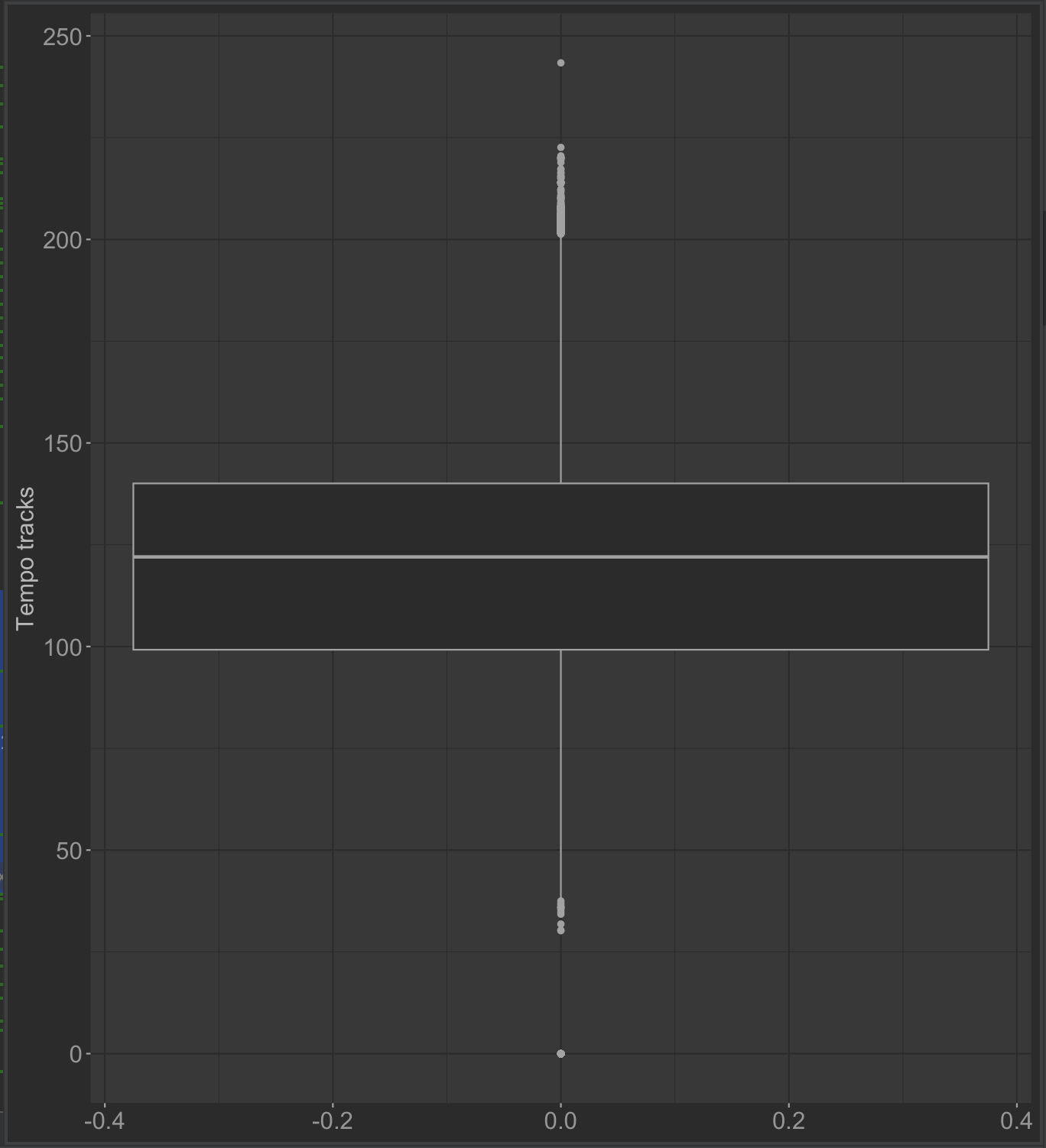
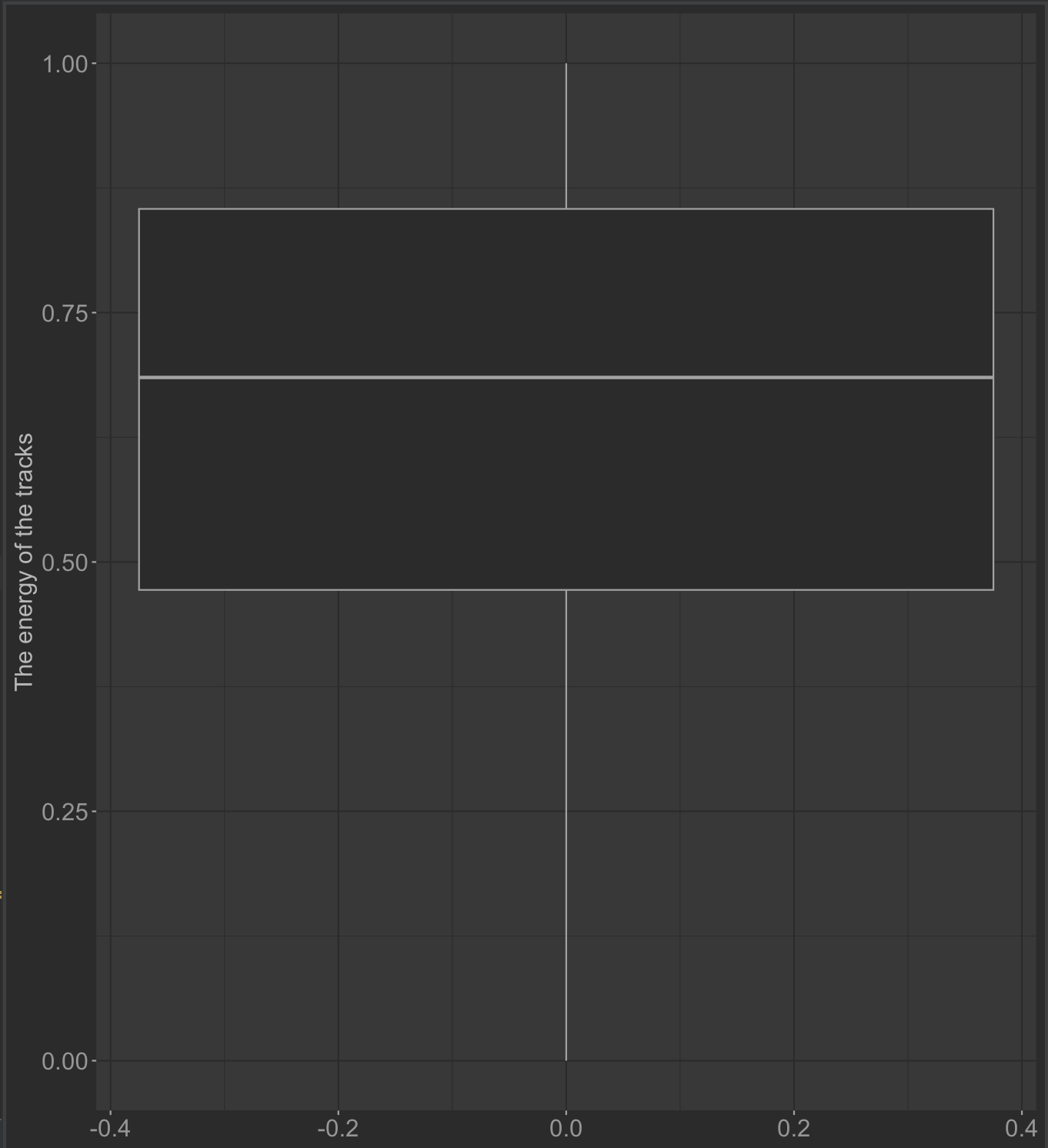
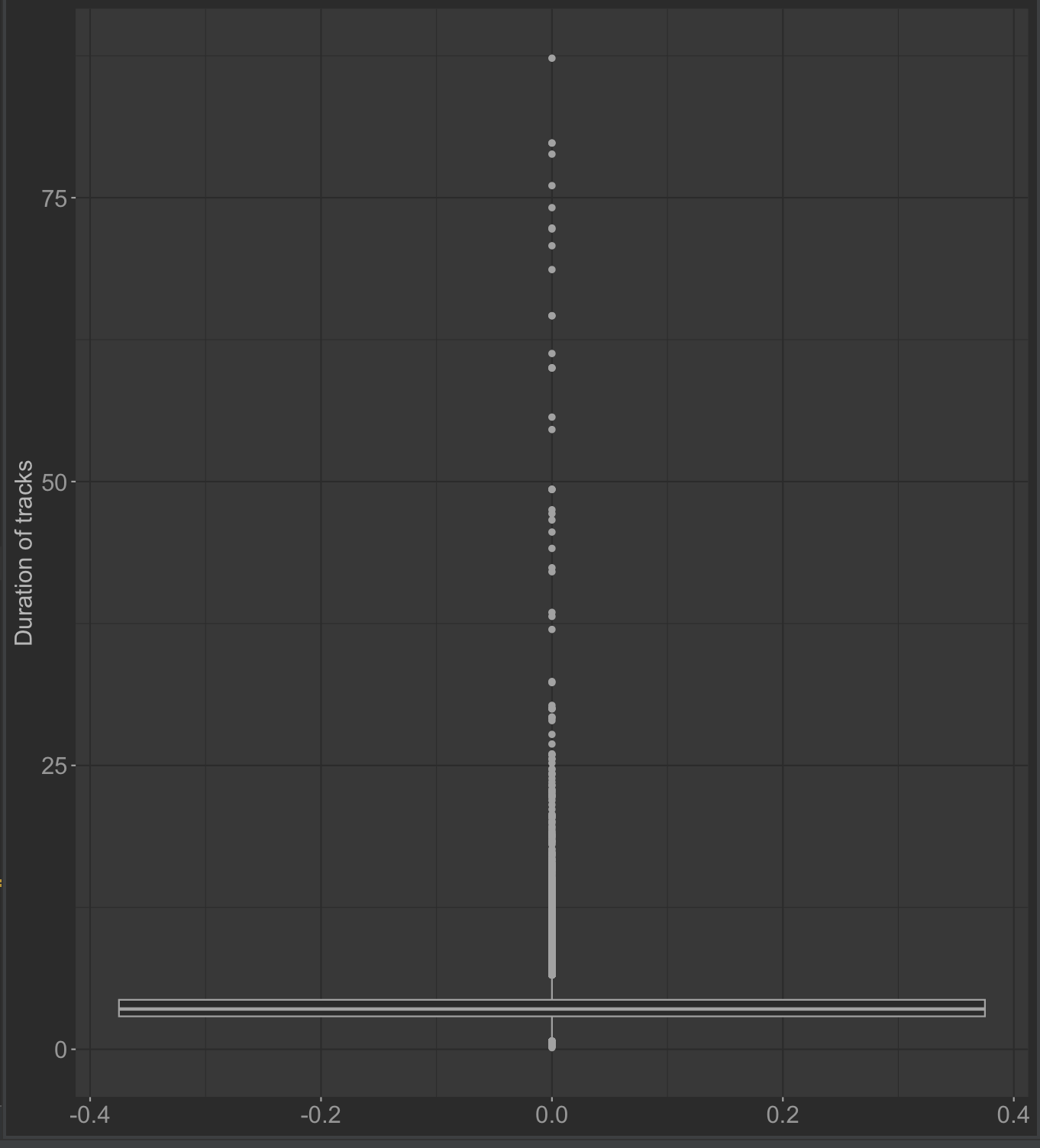
**Побудова boxplots**

Для візуального представлення розподілу кожної змінної та виявлення будь-яких викидів або екстремальних значень було створено функцію box\_plot.

Функція box\_plot створює графік для заданої змінної aes\_y та надає графіку заголовок title\_y. Функція також дозволяє користувачеві налаштувати розмір шрифту заголовка діаграми та підписів осей за допомогою параметрів title\_size і text\_size відповідно.

Функція використовує пакет ggplot2 з мови R для створення графіка у вигляді рамки. Параметр aes\_y задає змінну, яку потрібно відкласти на осі y діаграми. Функція викликається кілька разів з різними змінними для побудови відповідних діаграм, таких як popularity, valence, duration\_m, energy, tempo і speechiness. Ці змінні пов'язані з такими характеристиками музичних треків, як популярність, валентність, тривалість, енергійність, темп і промовистість.

Результати роботи функції для тих змінних, які фігурують в поставлених питаннях.

Висновки: Як ми бачимо, наближено до нормального розподілу є графік по змінній Valance. Майже всі інші розподіли зміщені та мають багато викидів.

**Перевірка викидів за допомогою функції Гампеля**

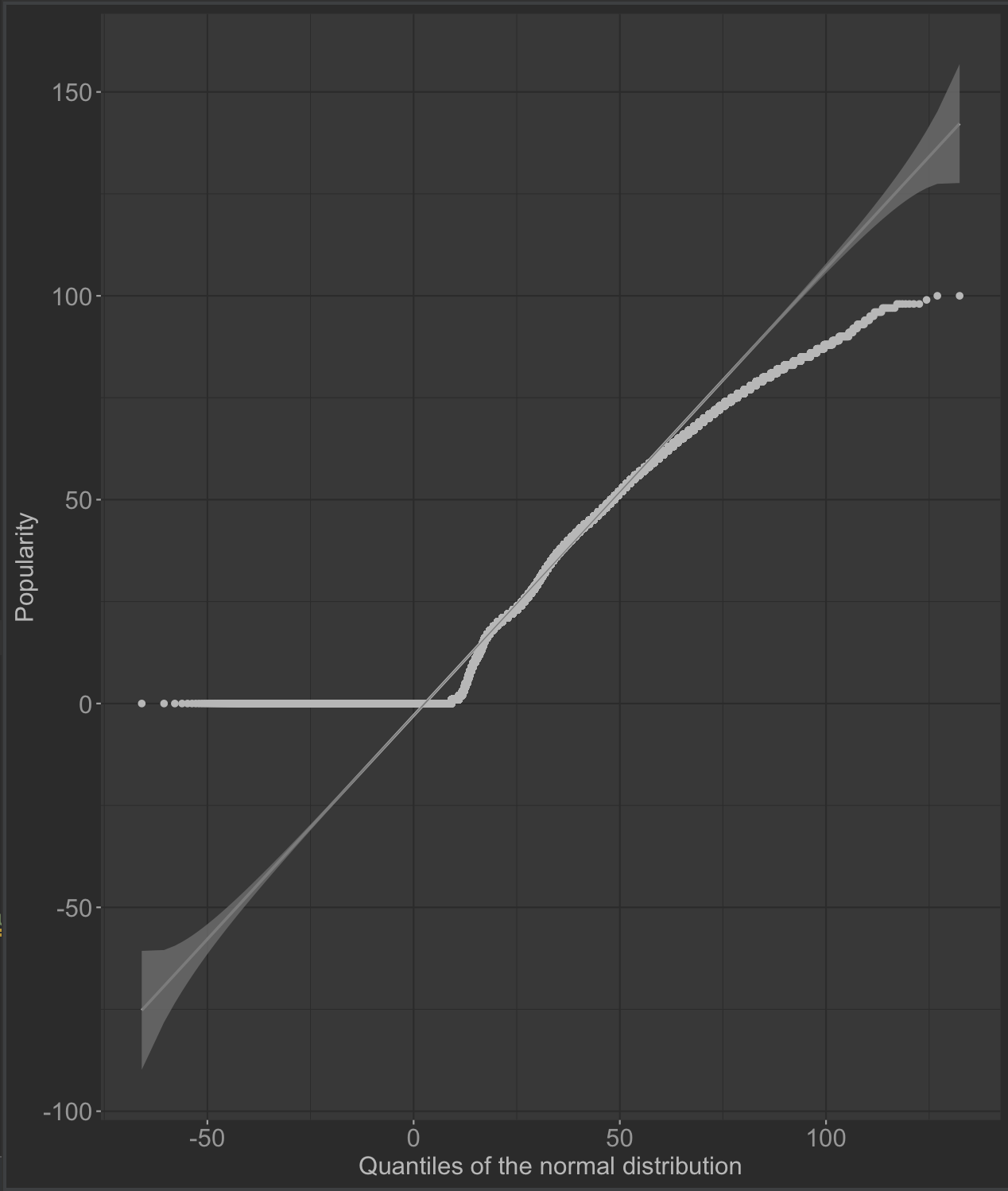
Для виявлення потенційних викидів у наборі даних, ми використали функцію Hampel\_filter. Вона перевіряє, чи є в стовпчику даних значення, що більш ніж в 3 рази перевищують медіанне абсолютне відхилення (MAD) від медіанного значення стовпчика. Якщо так, то ці значення позначаються як потенційні викиди.

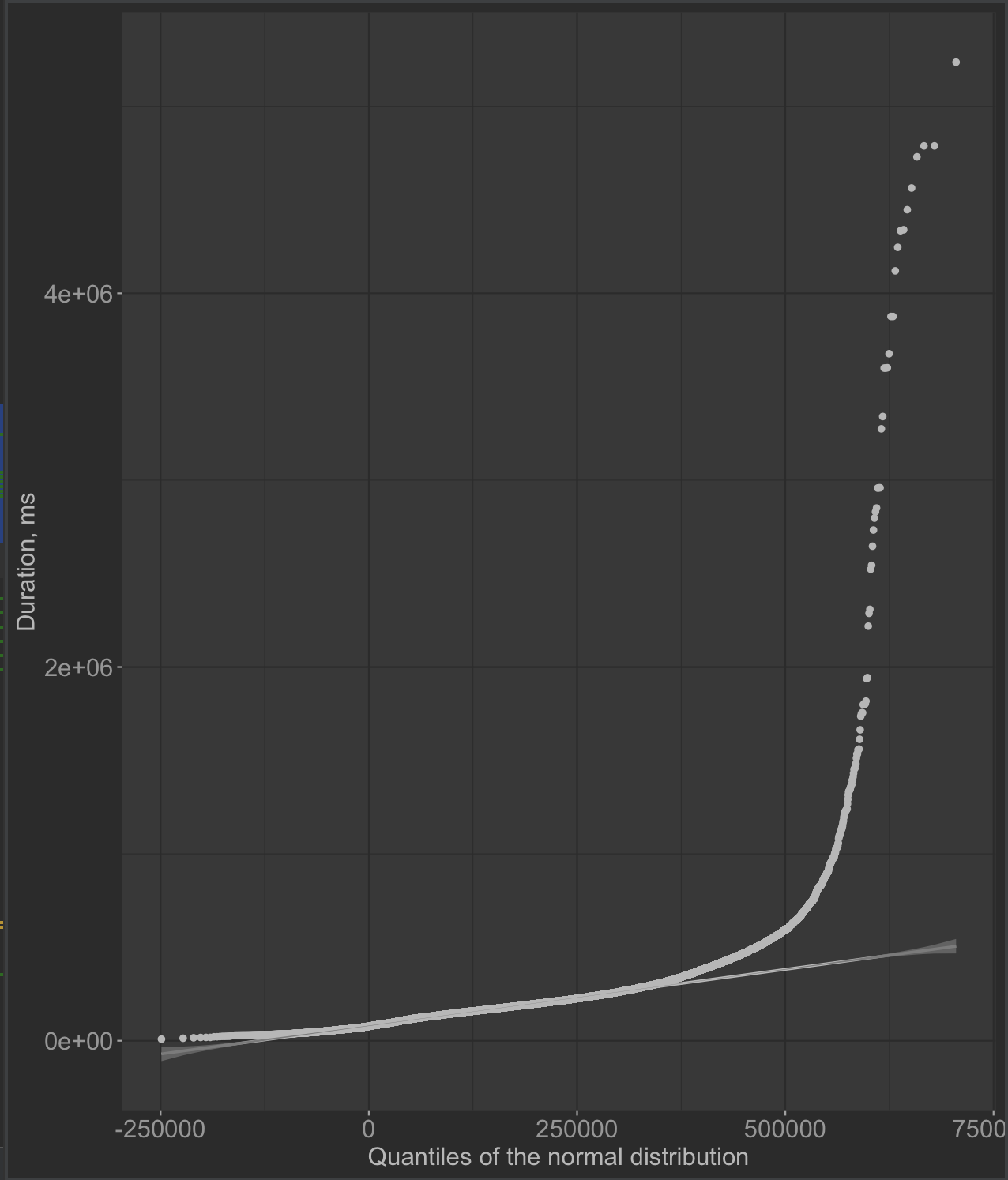
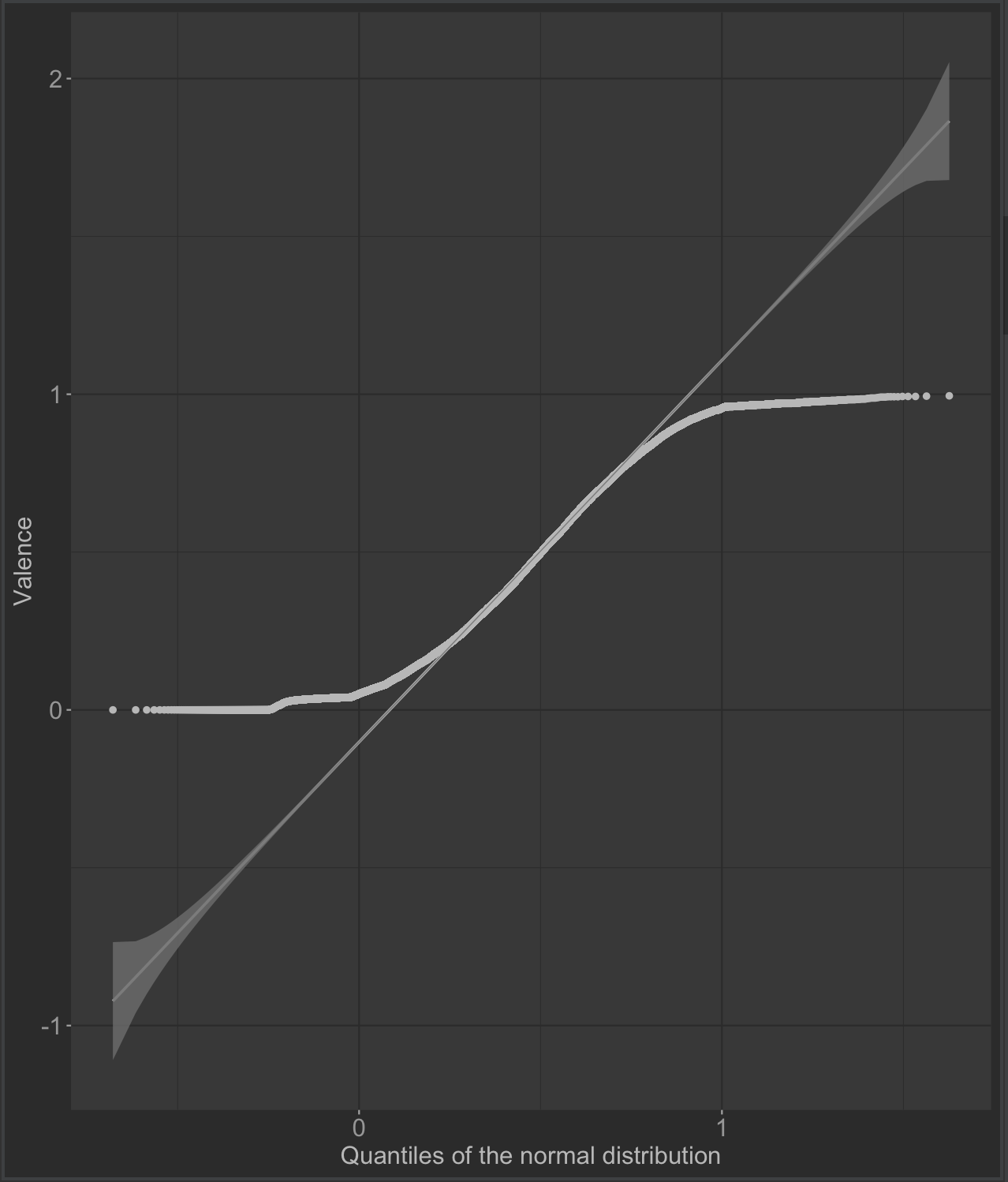
У нашому коді функція Hampel\_filter застосовується до кількох різних стовпців набору даних для виявлення потенційних викидів у кожному з них. Отримані відфільтровані набори даних потім роздруковуються і розташовуються в порядку спадання значення стовпчика, що перевіряється. Це може допомогти визначити, які конкретні значення в кожному стовпчику можуть бути викидами і потребують подальшого дослідження.

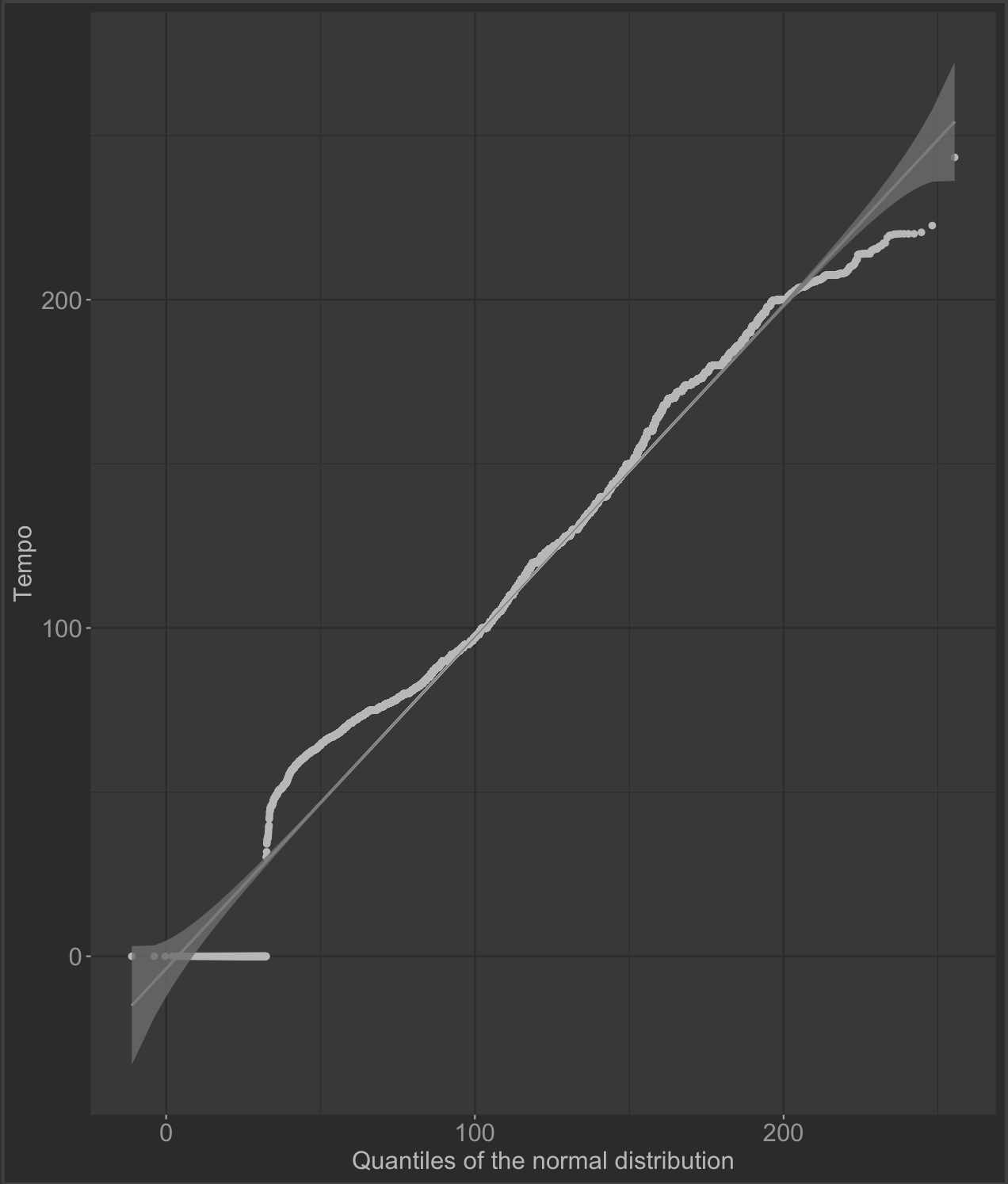
**Виведення QQ plots для змінних, що нас цікавлять**

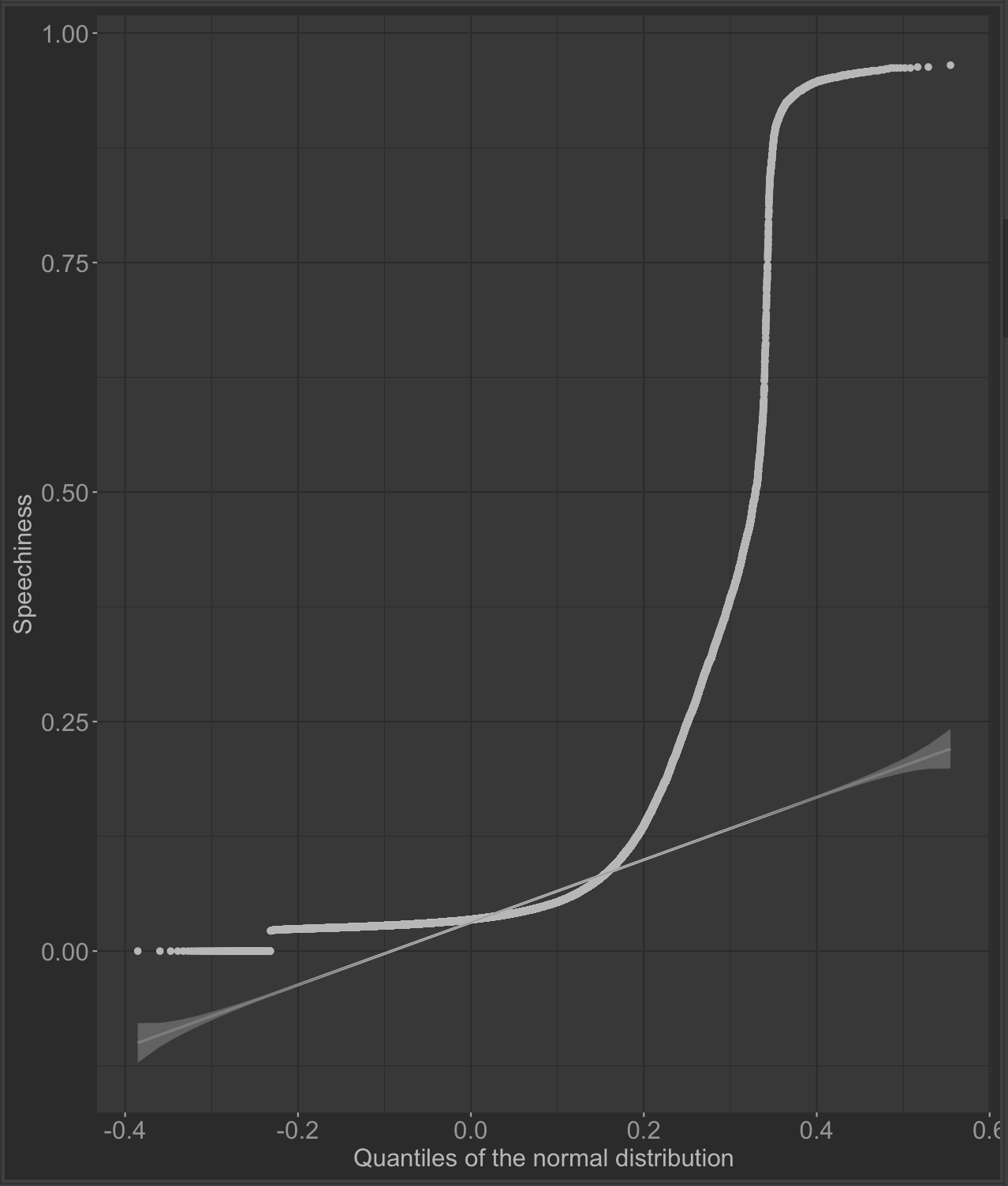
Функція qq\_plot\_builder використовується для створення QQ-графіків для змінних, що нас цікавлять. QQ-графіки - це спосіб візуально оцінити, чи є набір даних нормально розподіленим. Функція приймає два аргументи: aes, який вказує змінну, для якої будується графік, та y\_label, який є міткою для осі y. Функція використовує ggplot для створення графіка і включає три компоненти: точки, лінію і заштриховану смугу, які показують, наскільки близько дані відповідають нормальному розподілу. Функція застосовується до таких змінних: popularity, valence, duration\_ms, energy, tempo, and speechiness.

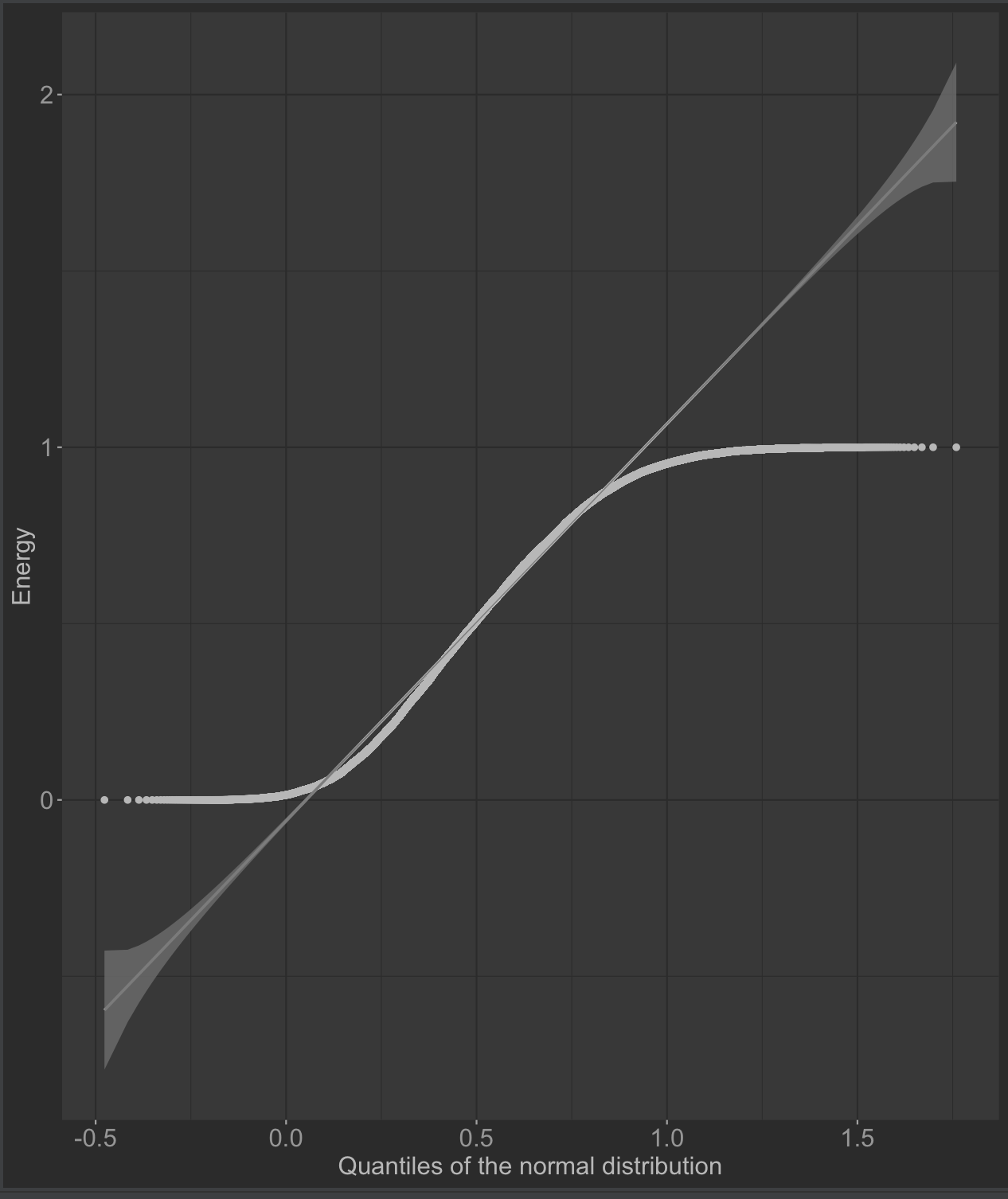
Результати роботи функції для тих змінних, які фігурують в поставлених питаннях:











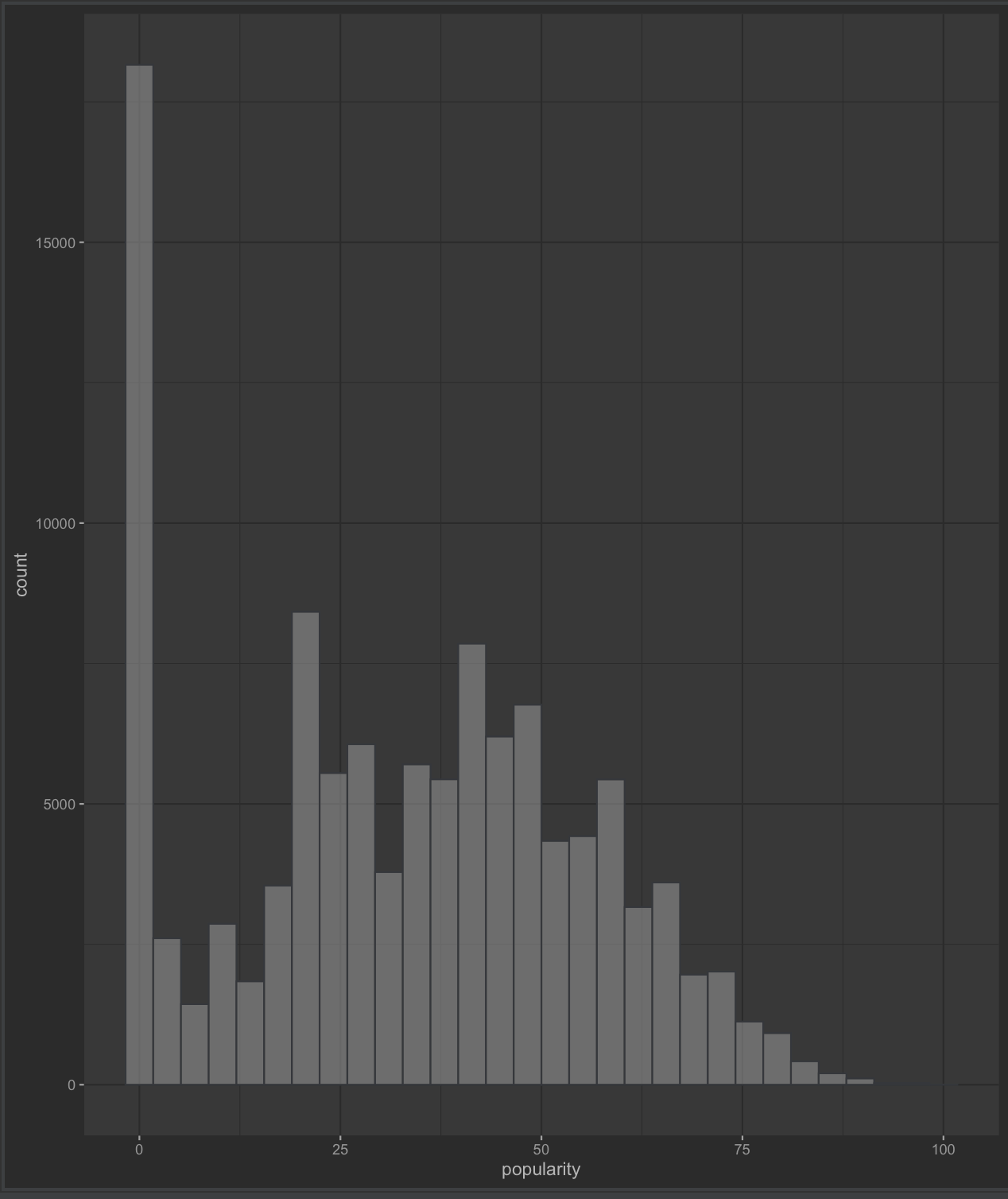
Висновки: По результатам QQ-графіків майже всі розподіли по всім змінним не є нормальними і мають досить багато викидів. Наближеним до нормального розподілу є графік по змінній tempo.

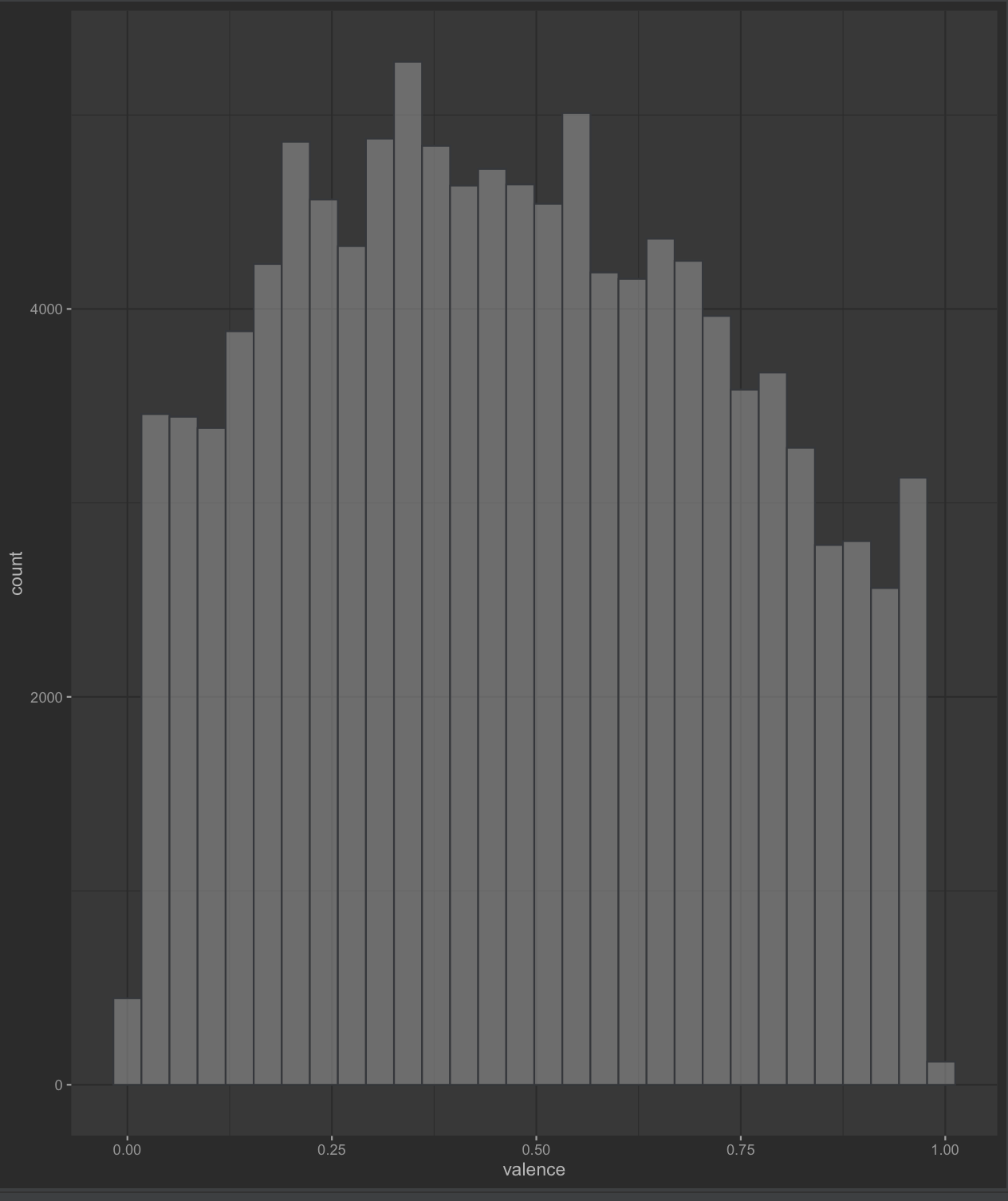
Дуже погана ситуація з speechiness і duration\_ms.

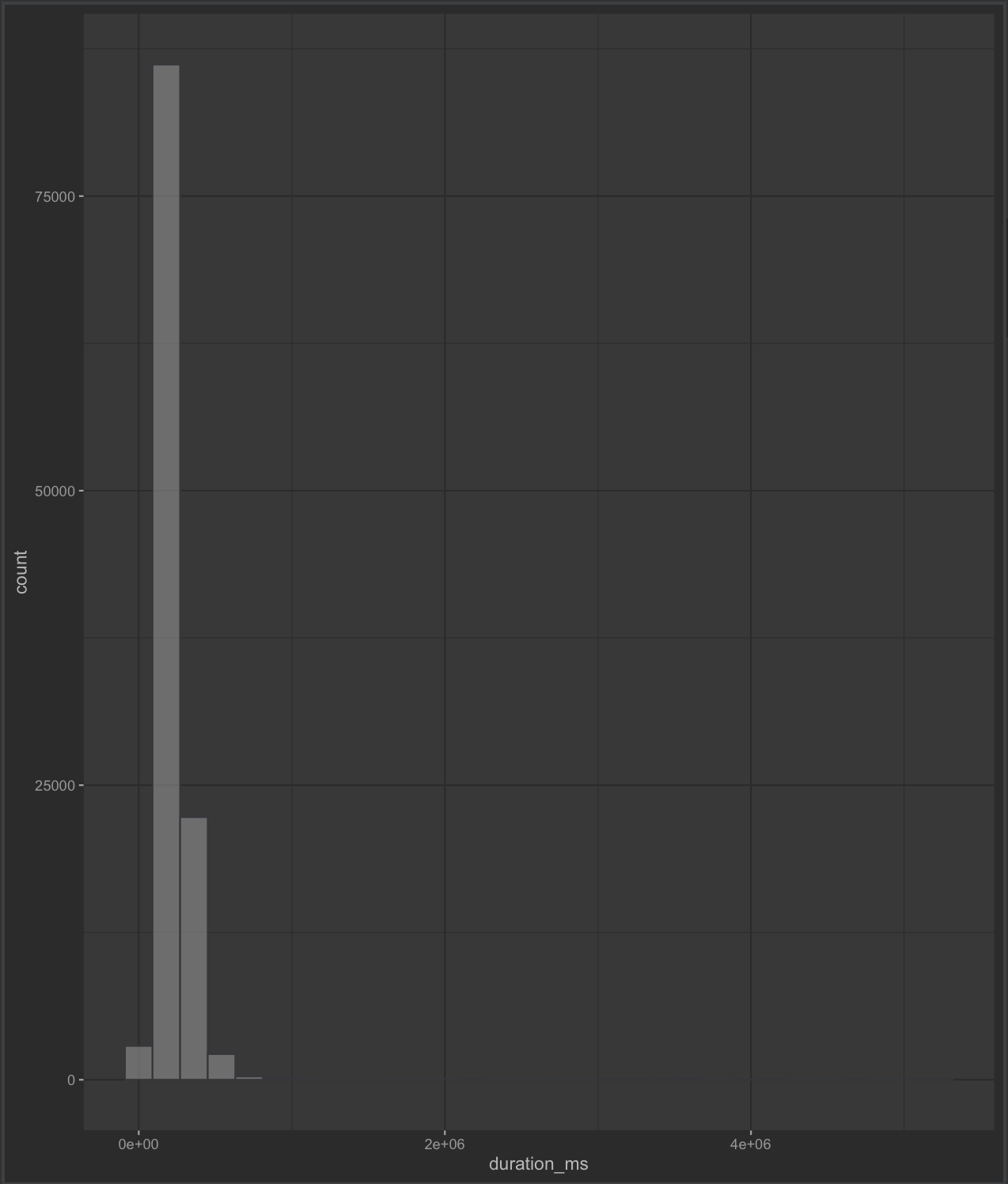
**Побудова гістограм для деяких неперервних змінних**

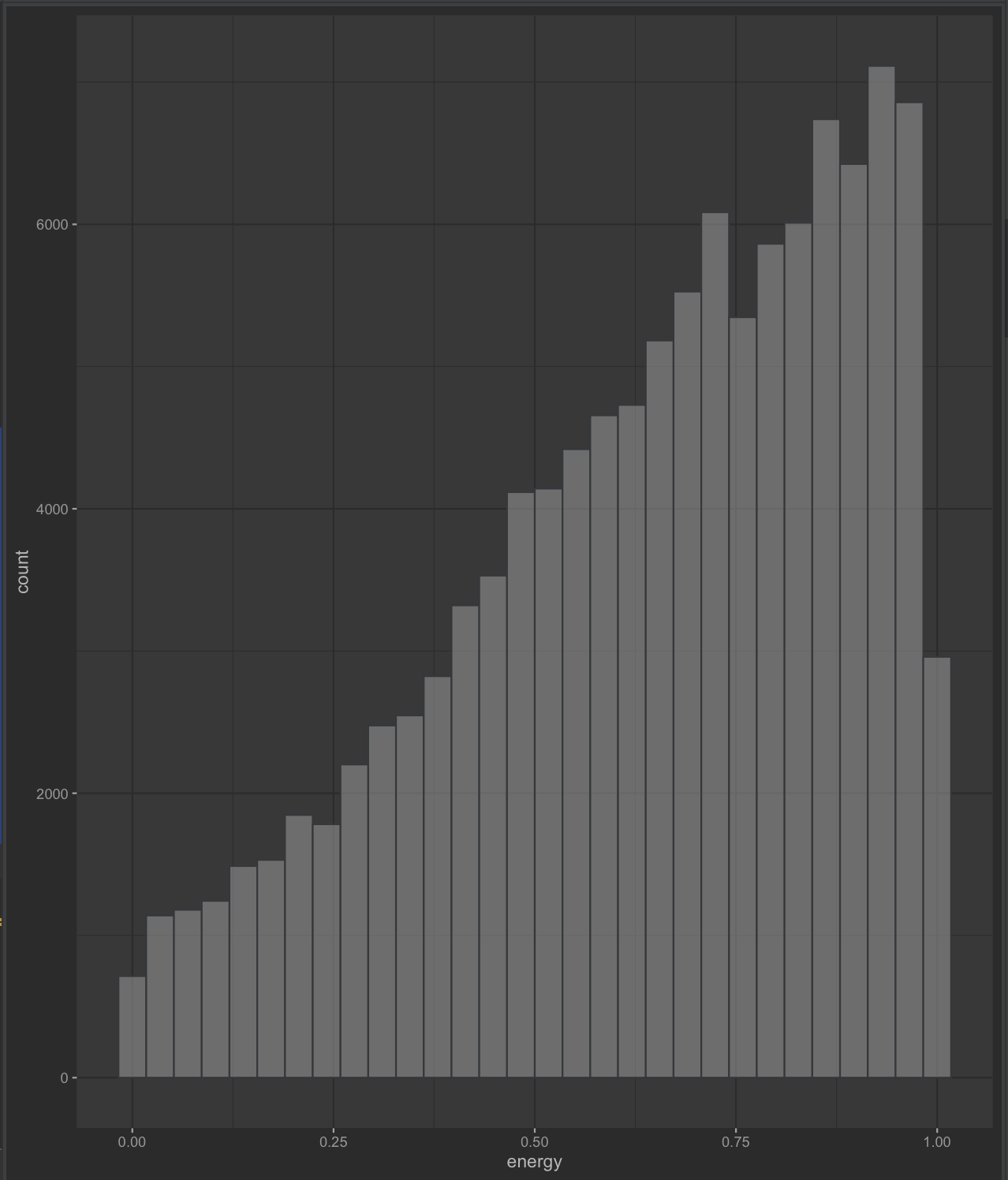
Функція histogram\_plot\_builder використовується для створення гістограм, які показують розподіл різних змінних у наборі даних. Гістограма - це графік, який показує, як часто певні значення з'являються у наборі даних. Функція приймає аргумент aes, який вказує, яку змінну використовувати для гістограми. Для кожної змінної функція використовує функцію ggplot() для створення об'єкта графіка і додає шар гістограми за допомогою функції geom\_histogram(). Параметр color задає колір межі смуг, а параметр alpha - прозорість смуг. Функція застосовується до таких змінних: popularity, valence, duration\_ms, energy, tempo, and speechiness.

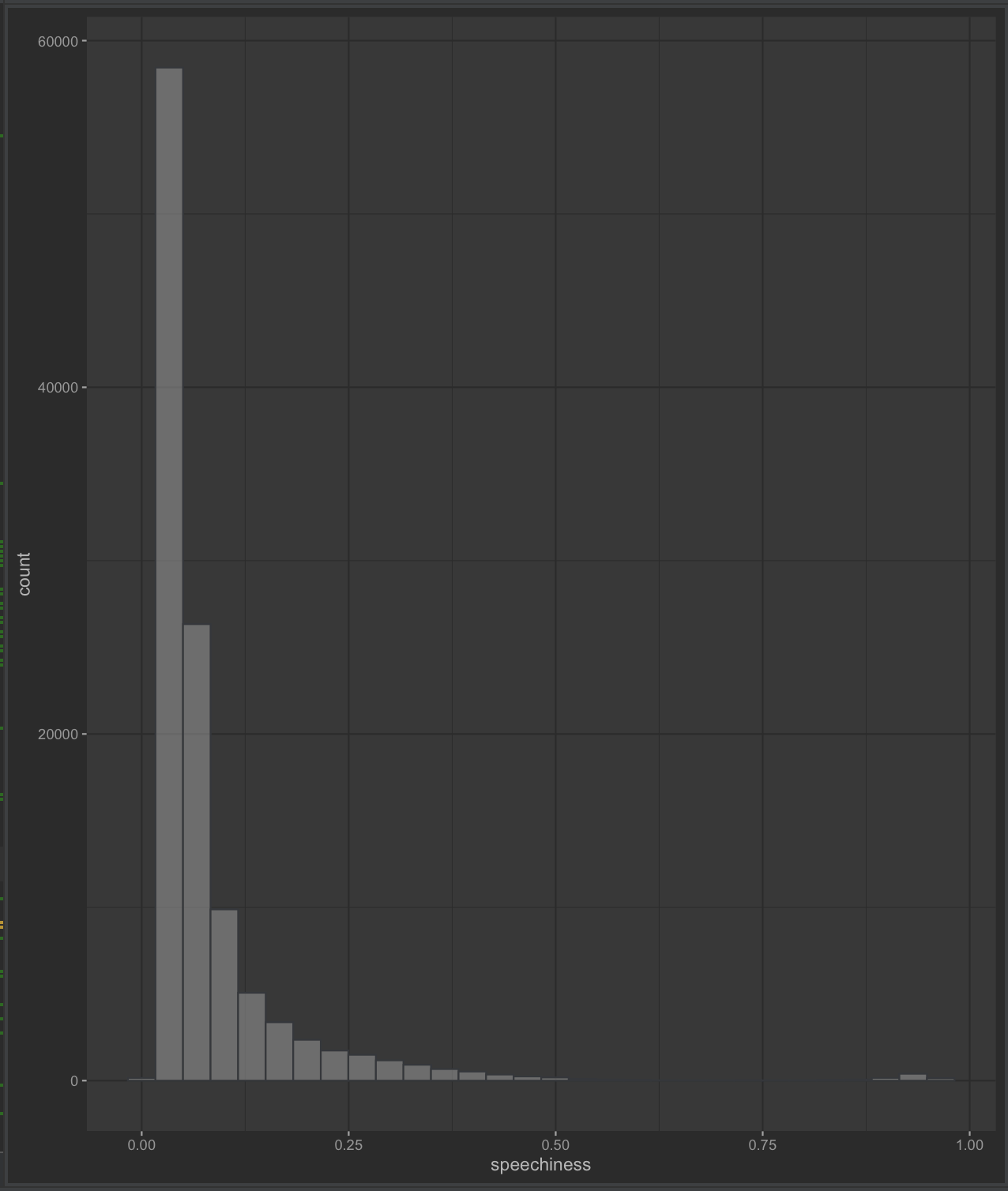
Отримані графіки показують розподіл змінної на осі х, а на осі у - частоту кожного значення. Смуги на графіку показують кількість разів, коли кожне значення з'являється в наборі даних.







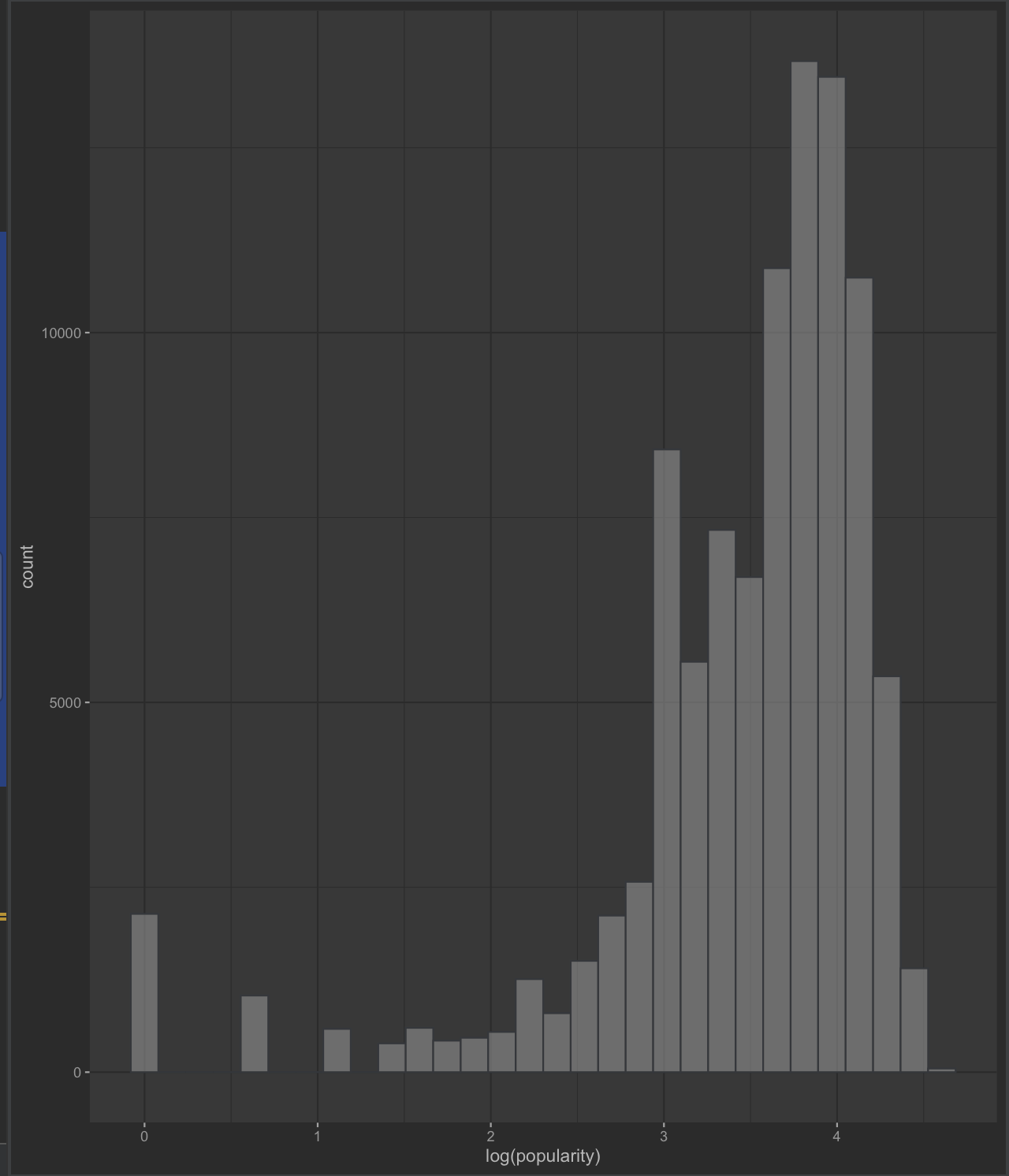


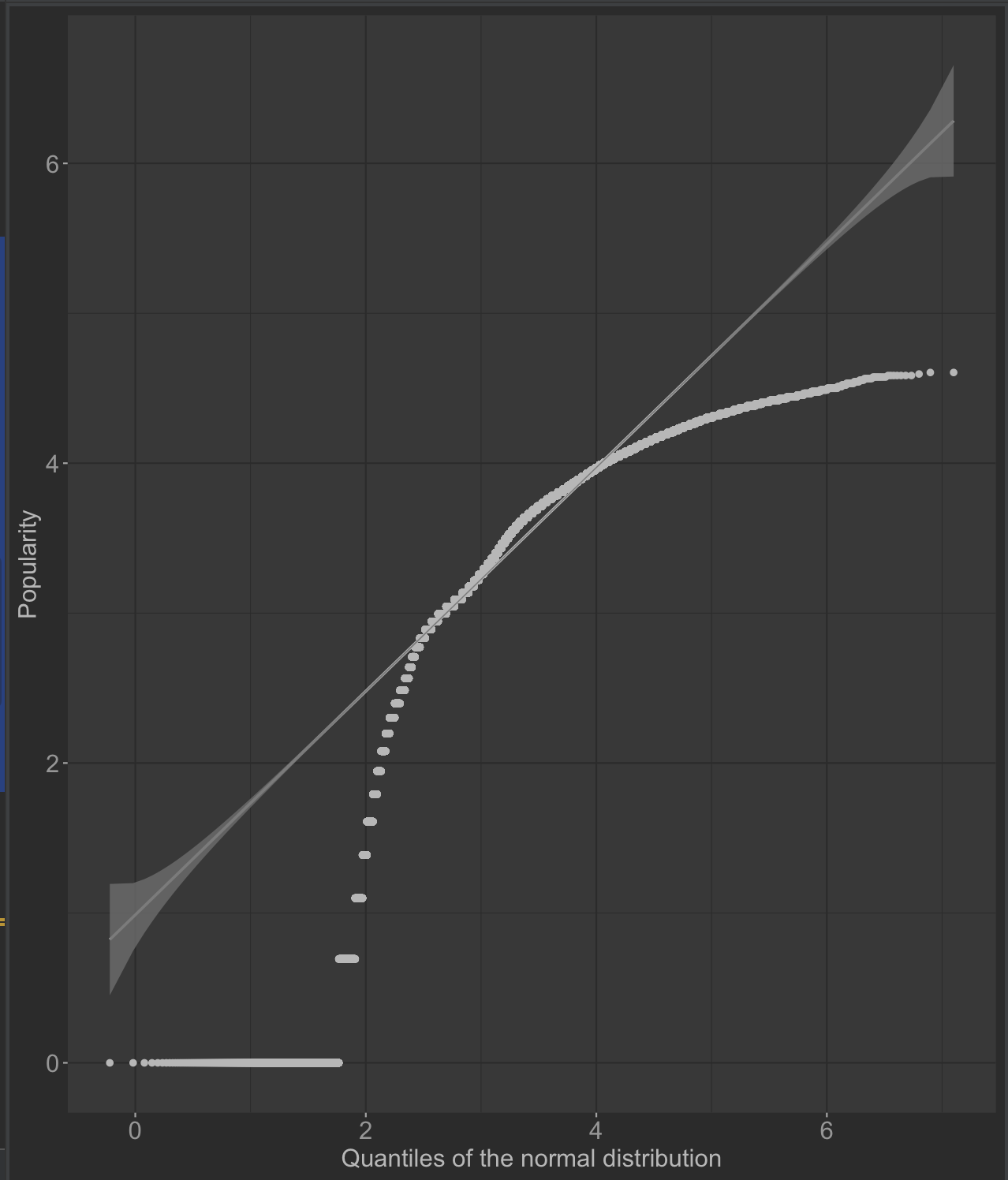


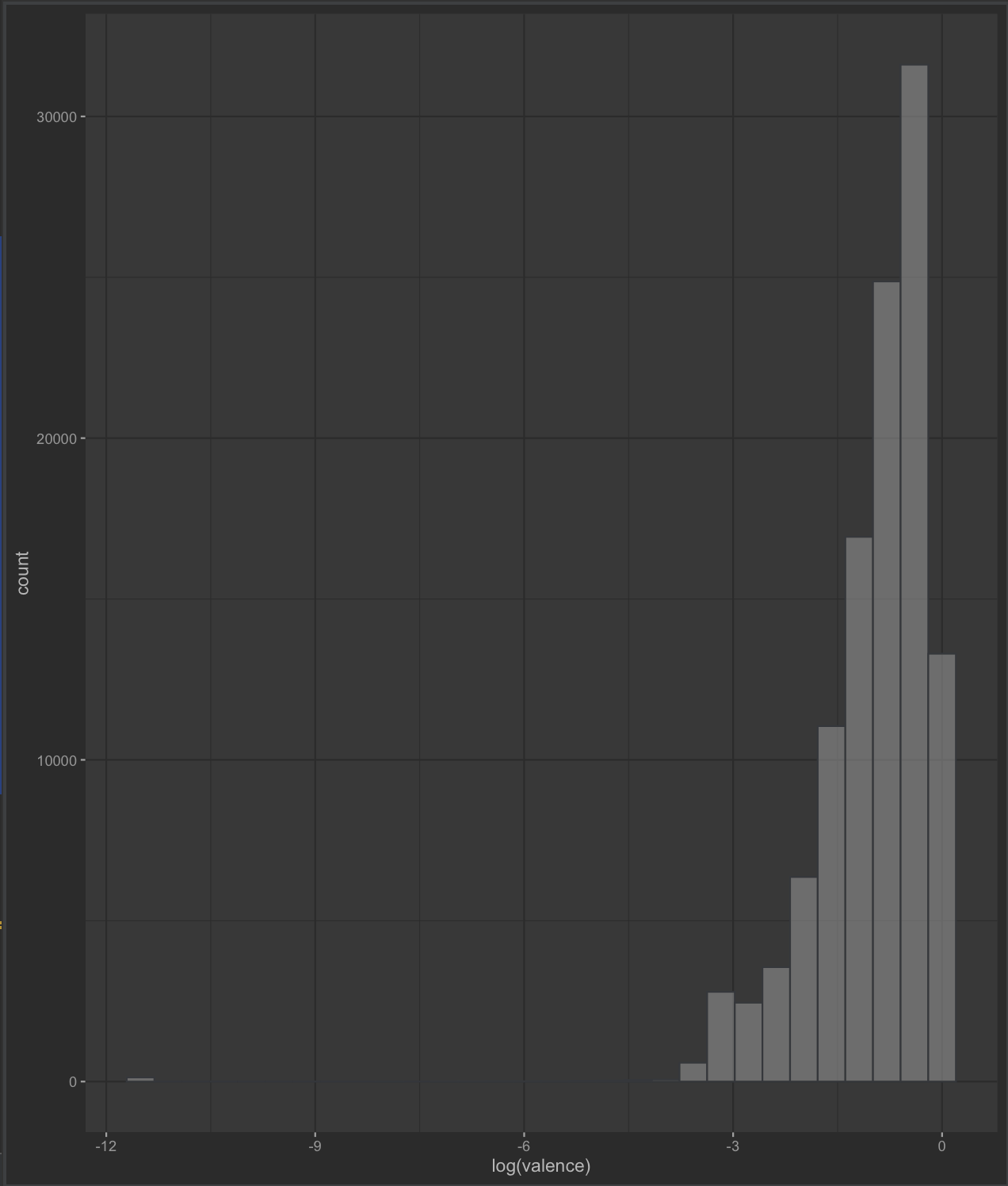
Висновки: Як ми бачимо, наближені до нормального розподілу є tempo і valance. Майже всі інші розподіли не є нормальними і мають досить багато викидів.

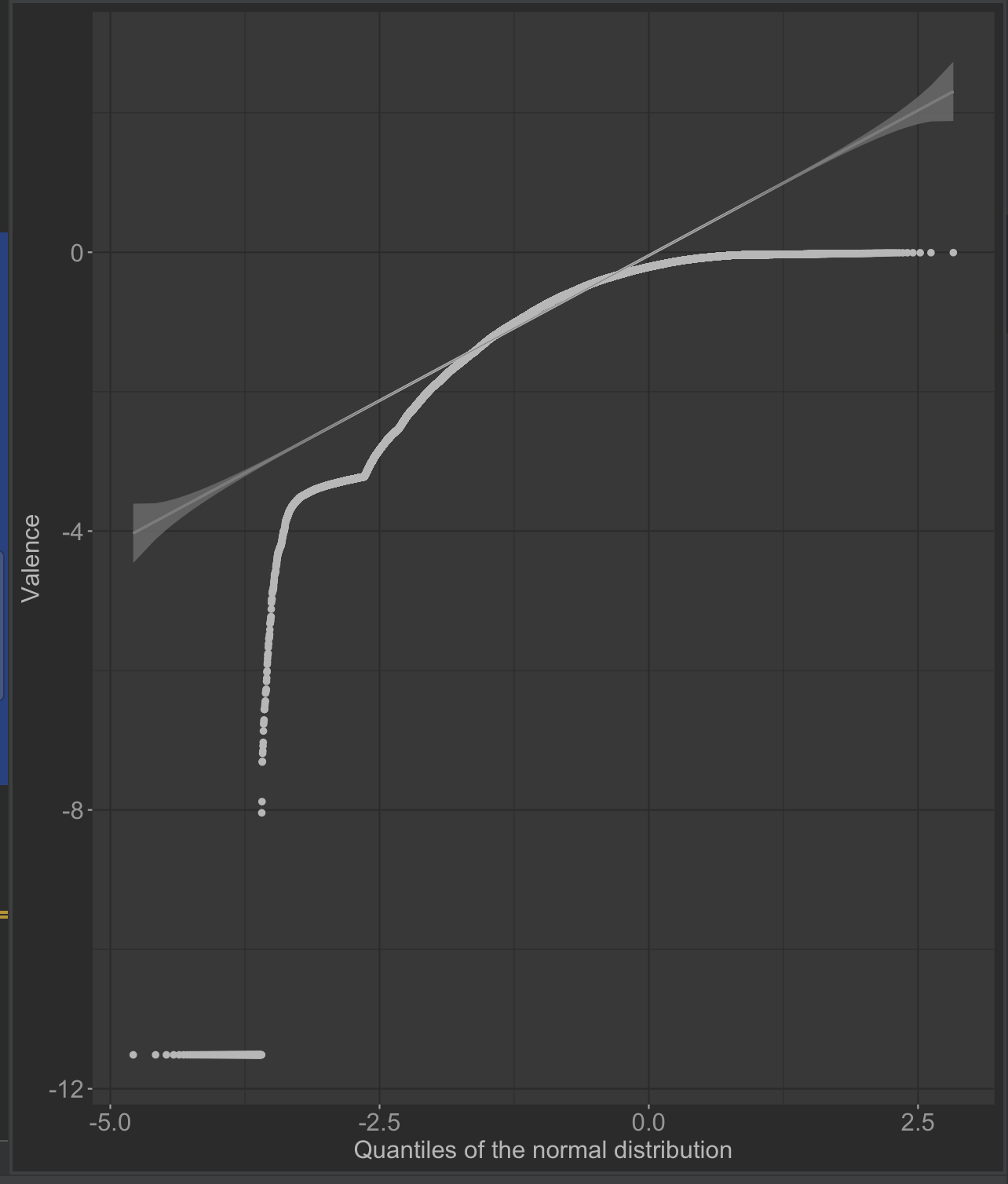
**Побудова гістограм та QQ-графіків прологаритмованих змінних**

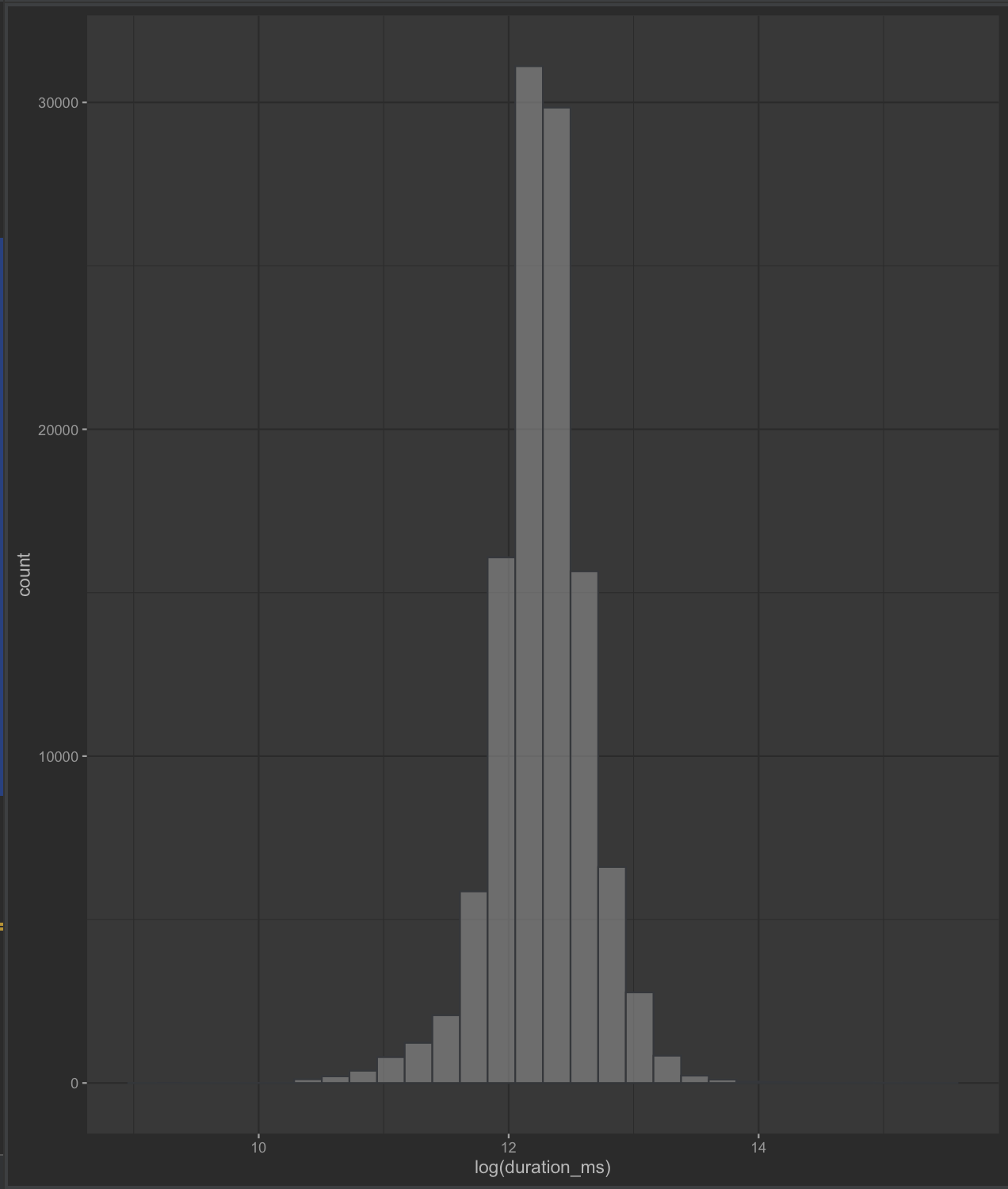
Для побудови гістограм і QQ-графіків прологарифмічних змінних ми використовуємо вже написані функції histogram\_plot\_builder і qq\_plot\_builder, але застосовуємо їх до змінних після піднесення їх до логарифму. Це допомагає перетворити асиметричні дані в більш нормально розподілену форму, яка краще підходить для статистичного аналізу. Ми створюємо ці графіки для змінних popularity, valence, duration\_ms, energy, tempo, and speechiness. Для кожної змінної ми спочатку створюємо гістограму, а потім QQ-графік. Гістограма показує розподіл змінної, тоді як QQ-діаграма порівнює її розподіл з нормальним.

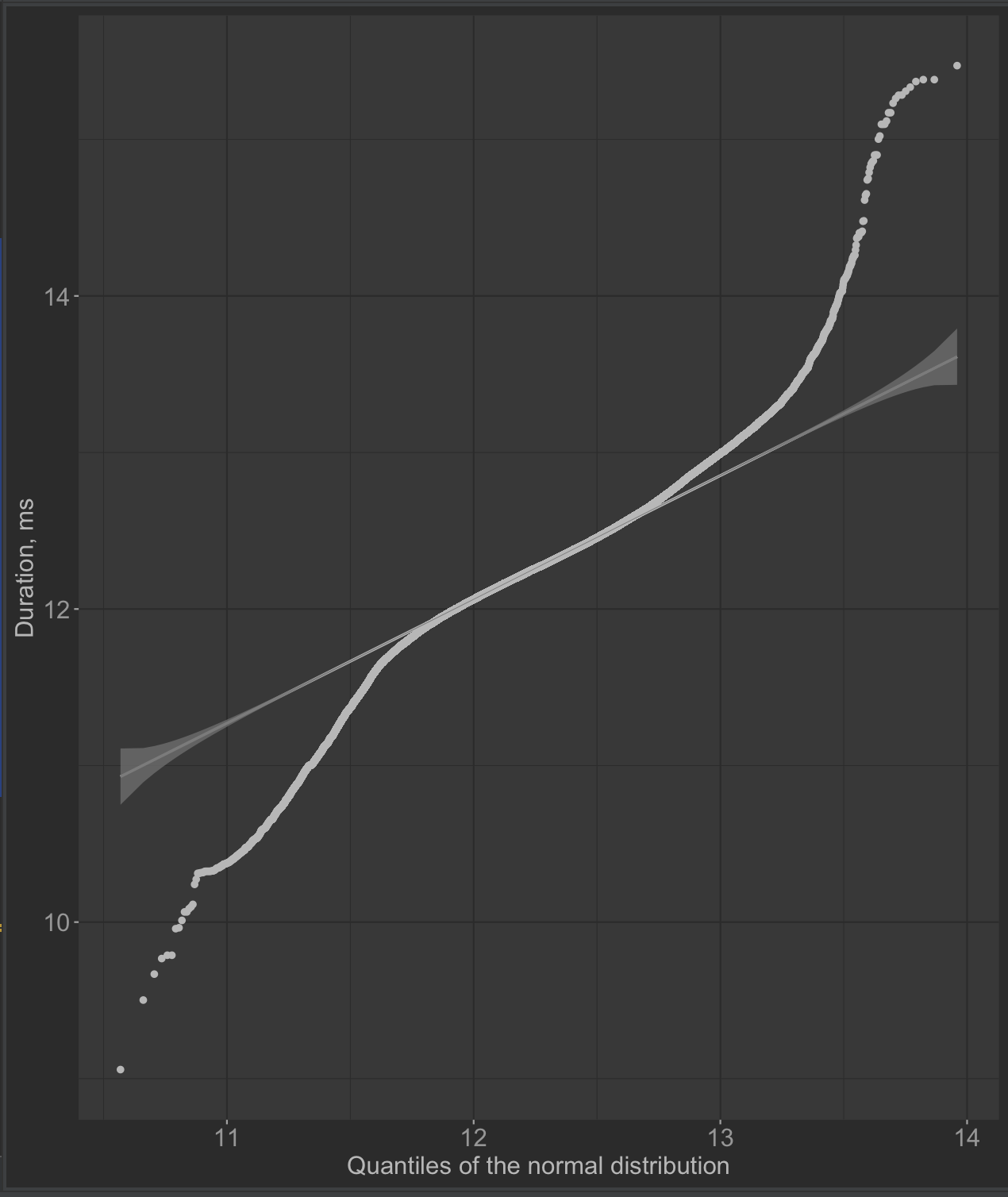


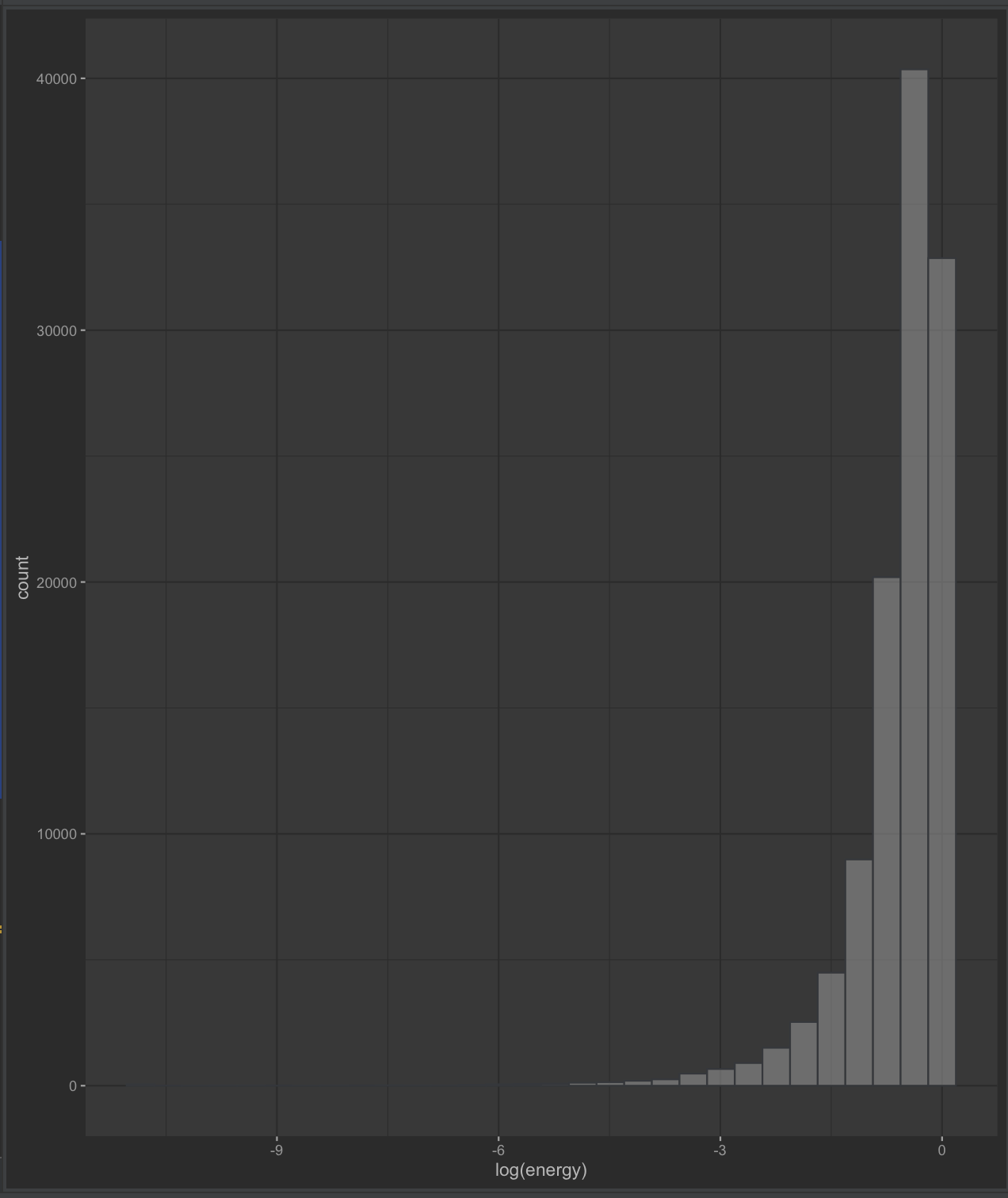


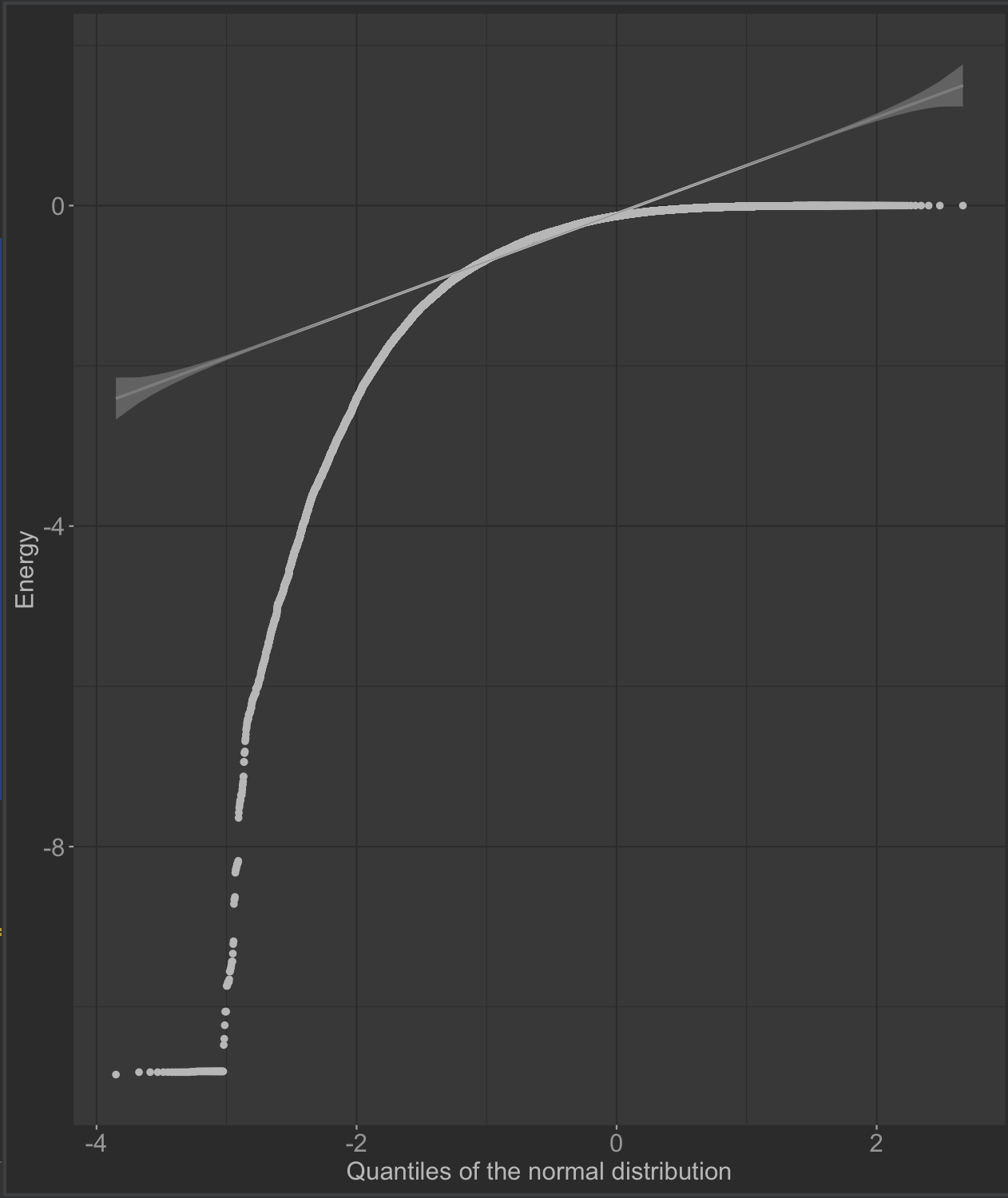


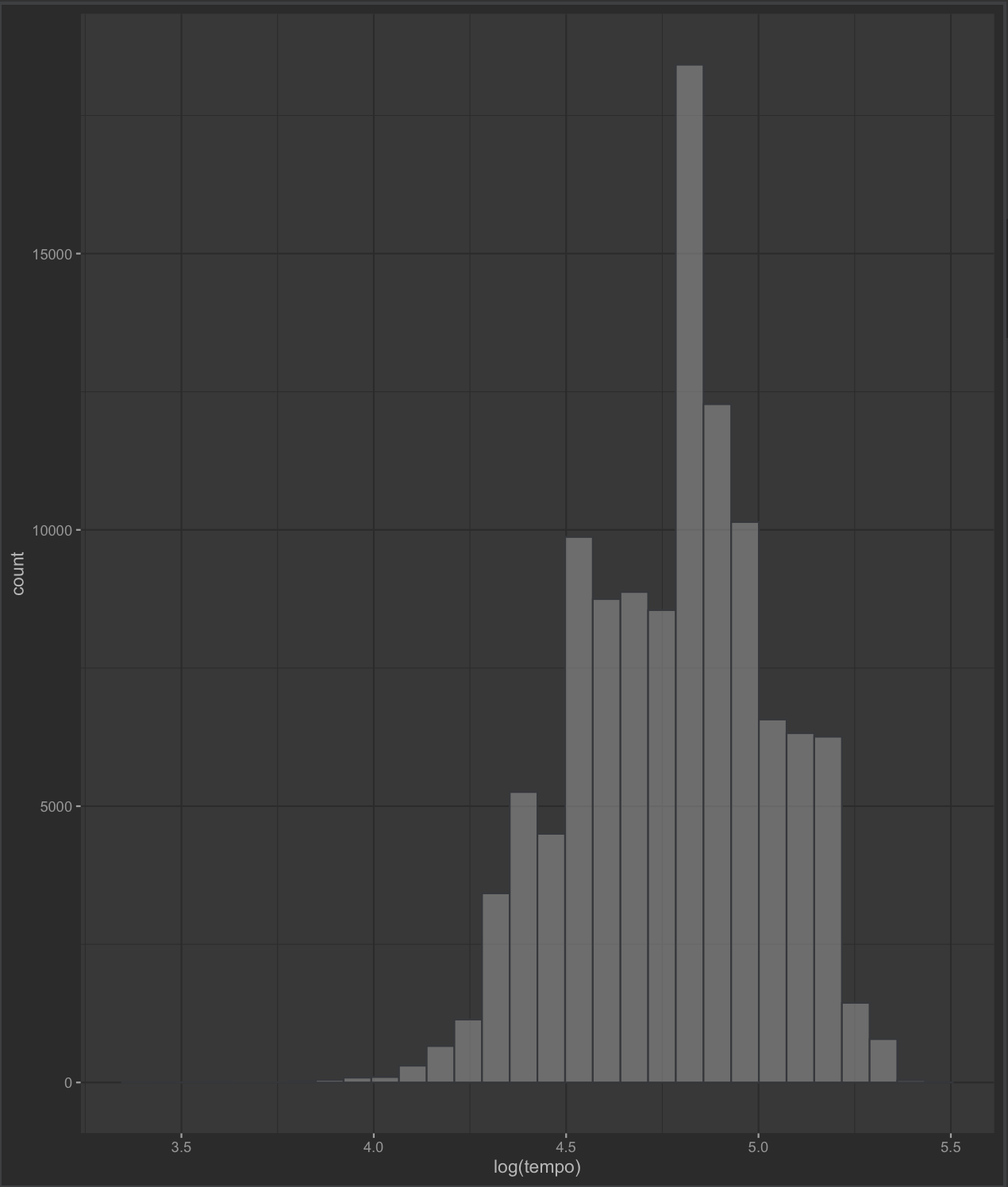


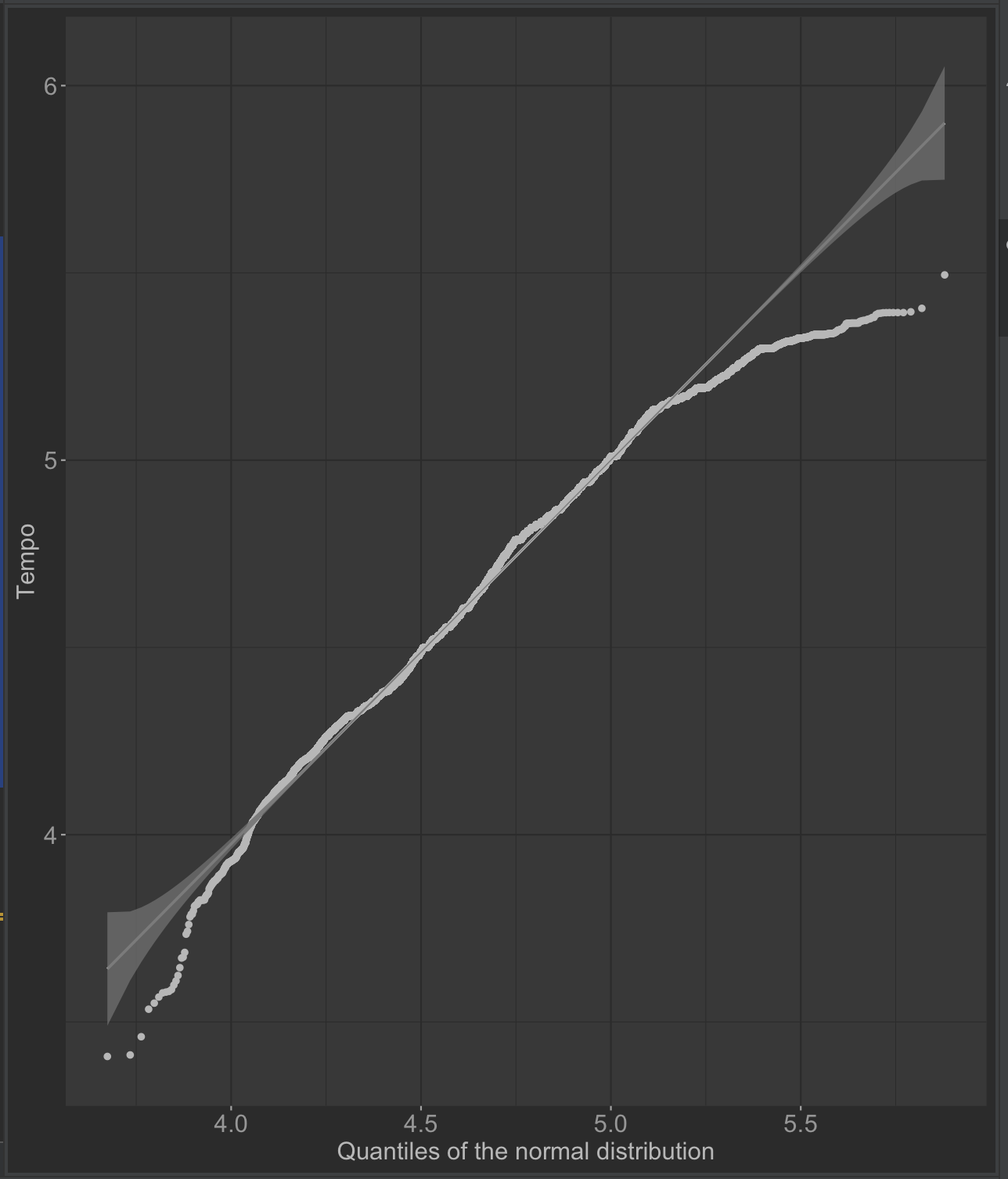


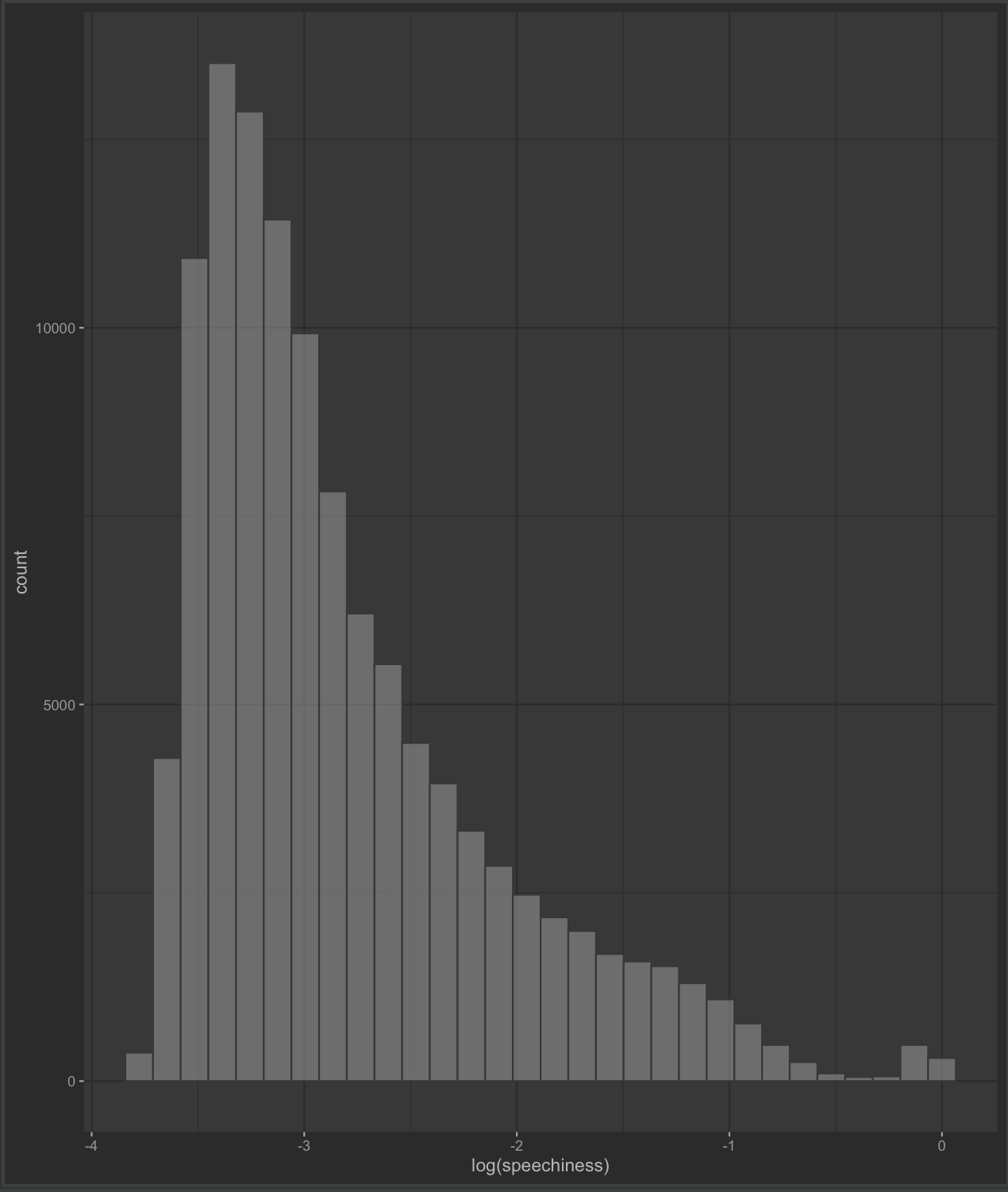


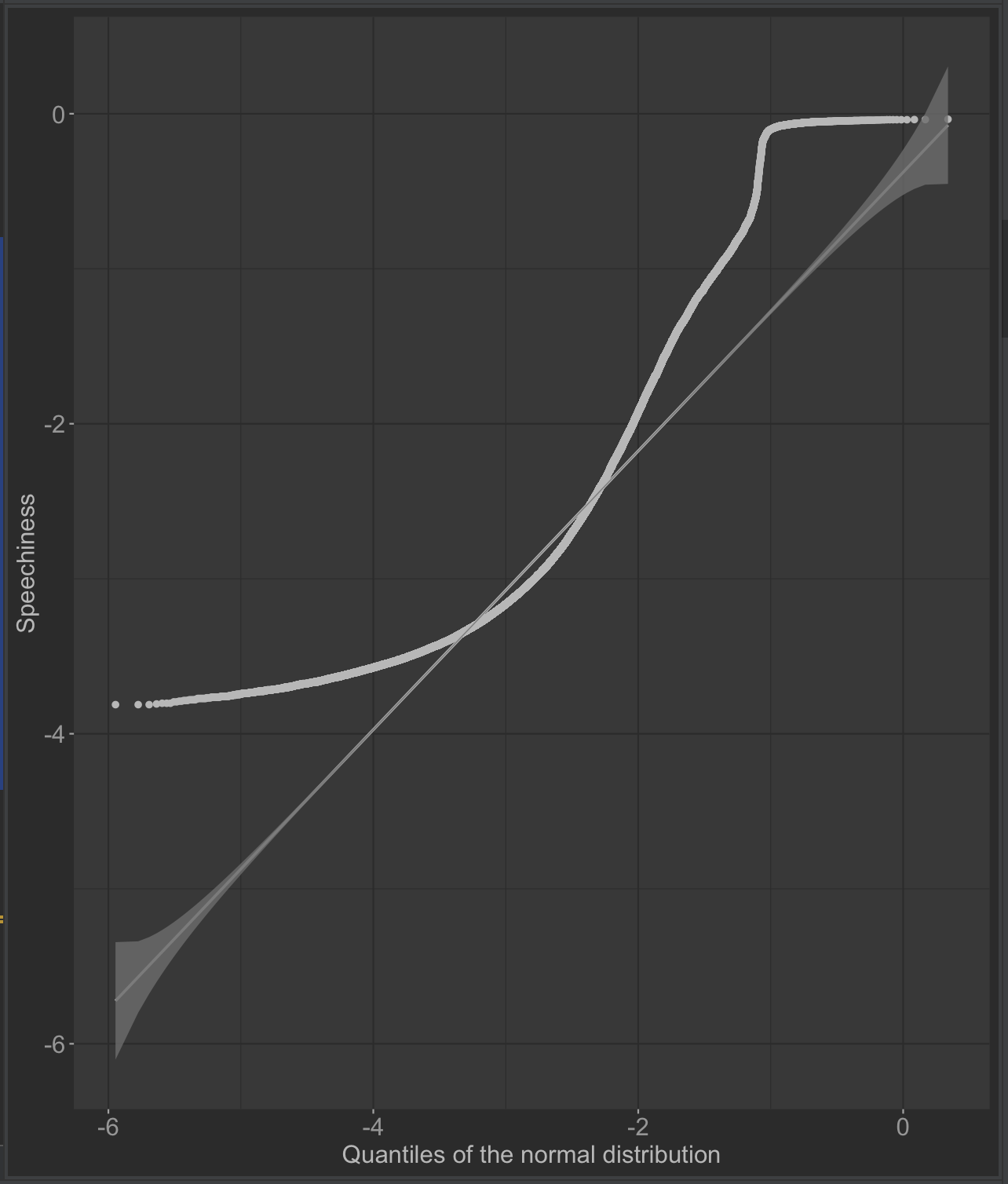










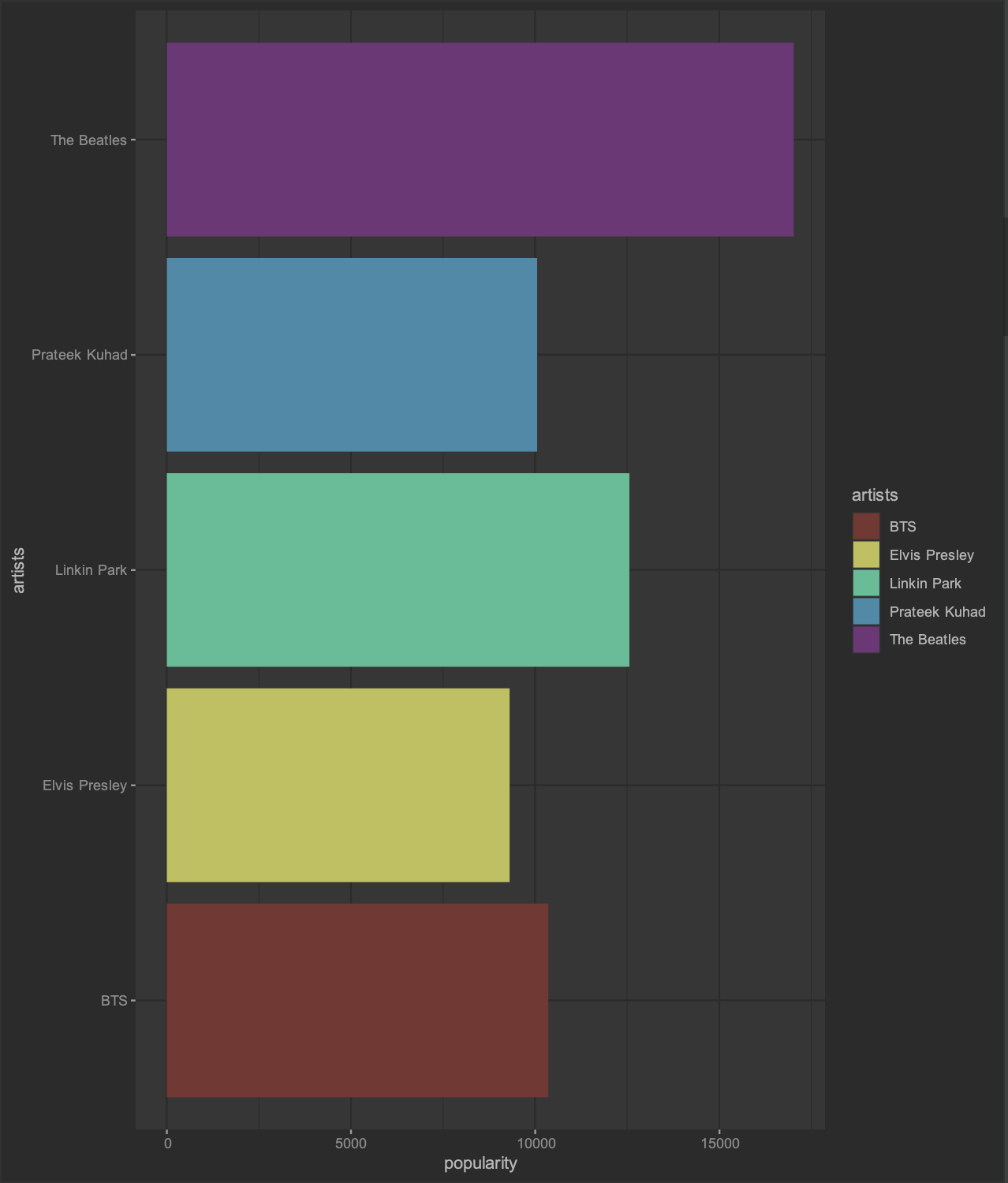


Висновки: На основі гістограм та QQ-діаграм прологарифмічних змінних можна зробити висновок, що піднесення змінних до логарифму не призвело до суттєвого покращення нормального розподілу даних. Насправді, в деяких випадках, наприклад, у випадку з popularity, піднесення до логарифму фактично призвело до погіршення розподілу. Тому, можливо, немає необхідності використовувати прологарифмічні перетворення для цих змінних у подальшому аналізі.

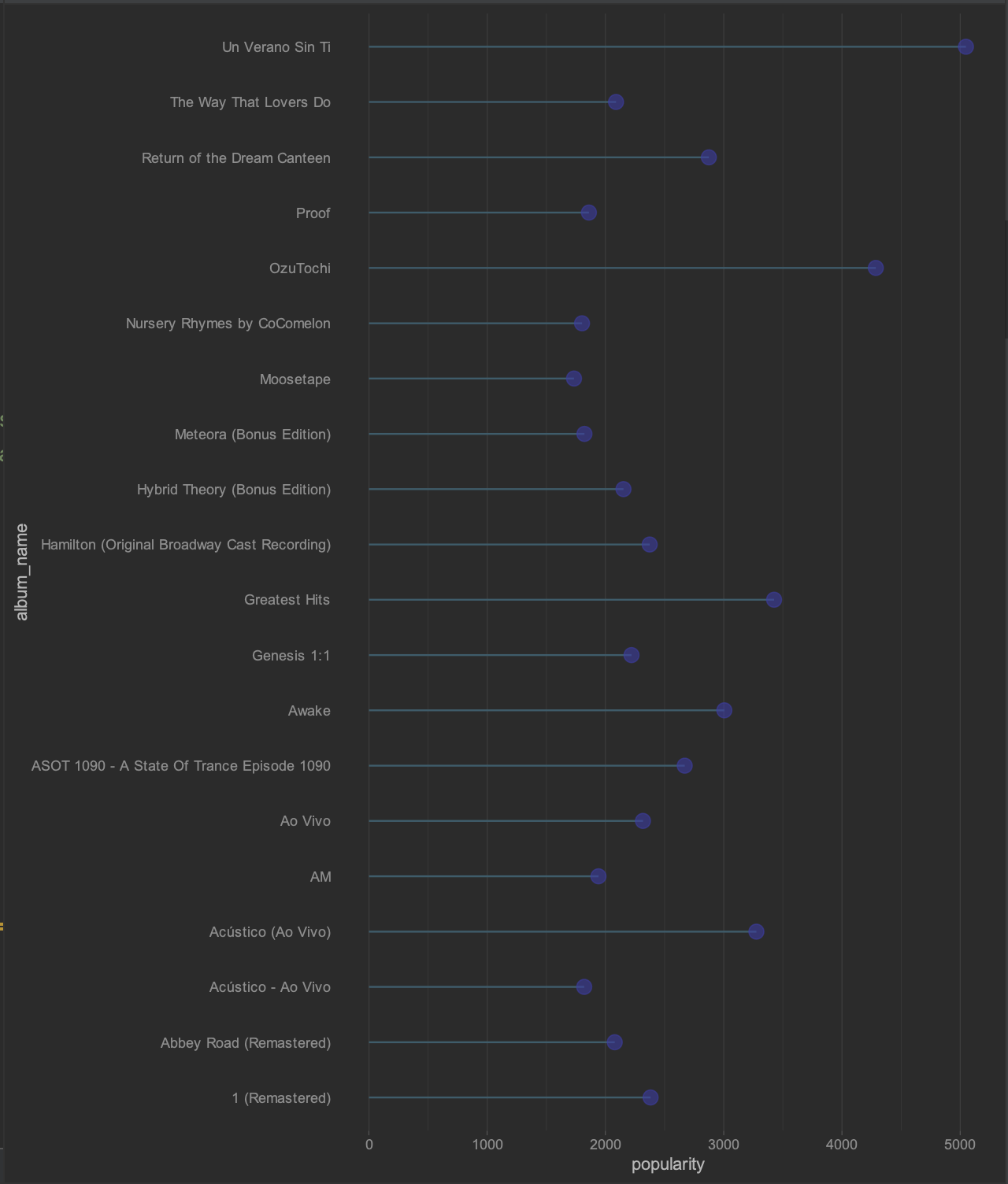
**Виявлення найбільш популярних об'єктів в залежності від певних харакеристик**

В цій частині ми знаходимо найпопулярніших виконавців, альбоми та жанри на основі їхньої загальної популярності. Потім він створюємо гістограми для візуального відображення цієї інформації.

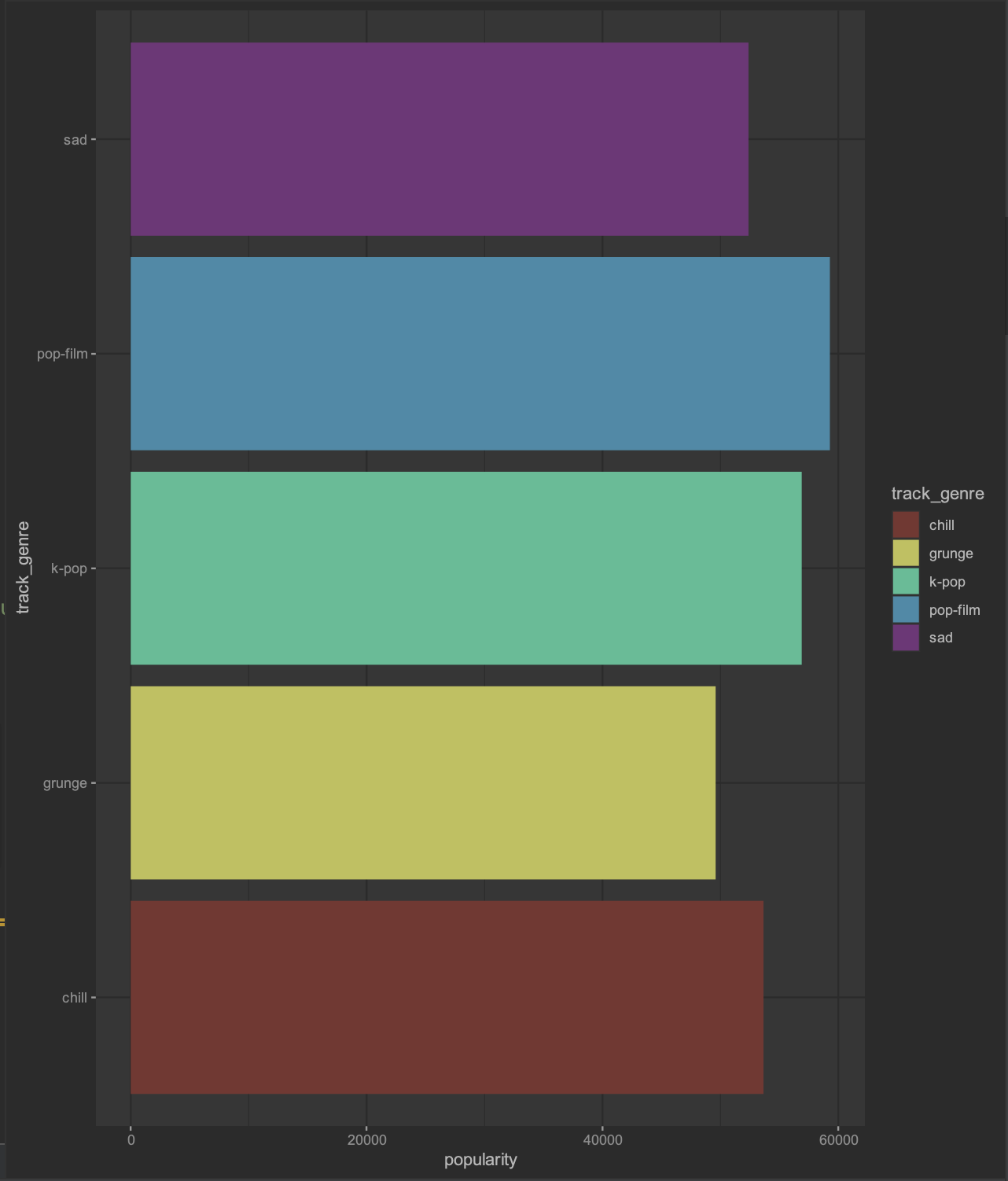
most\_popular\_artists використовує функцію group\_by для групування даних за виконавцями, а потім summarise для обчислення суми їхніх рейтингів популярності. Результуючий фрейм даних впорядковується за спаданням рейтингу популярності, а 5 найкращих виконавців вибираються за допомогою функції head. Потім код друкує отриманий фрейм даних і створює гістограму, використовуючи ggplot для візуалізації даних.



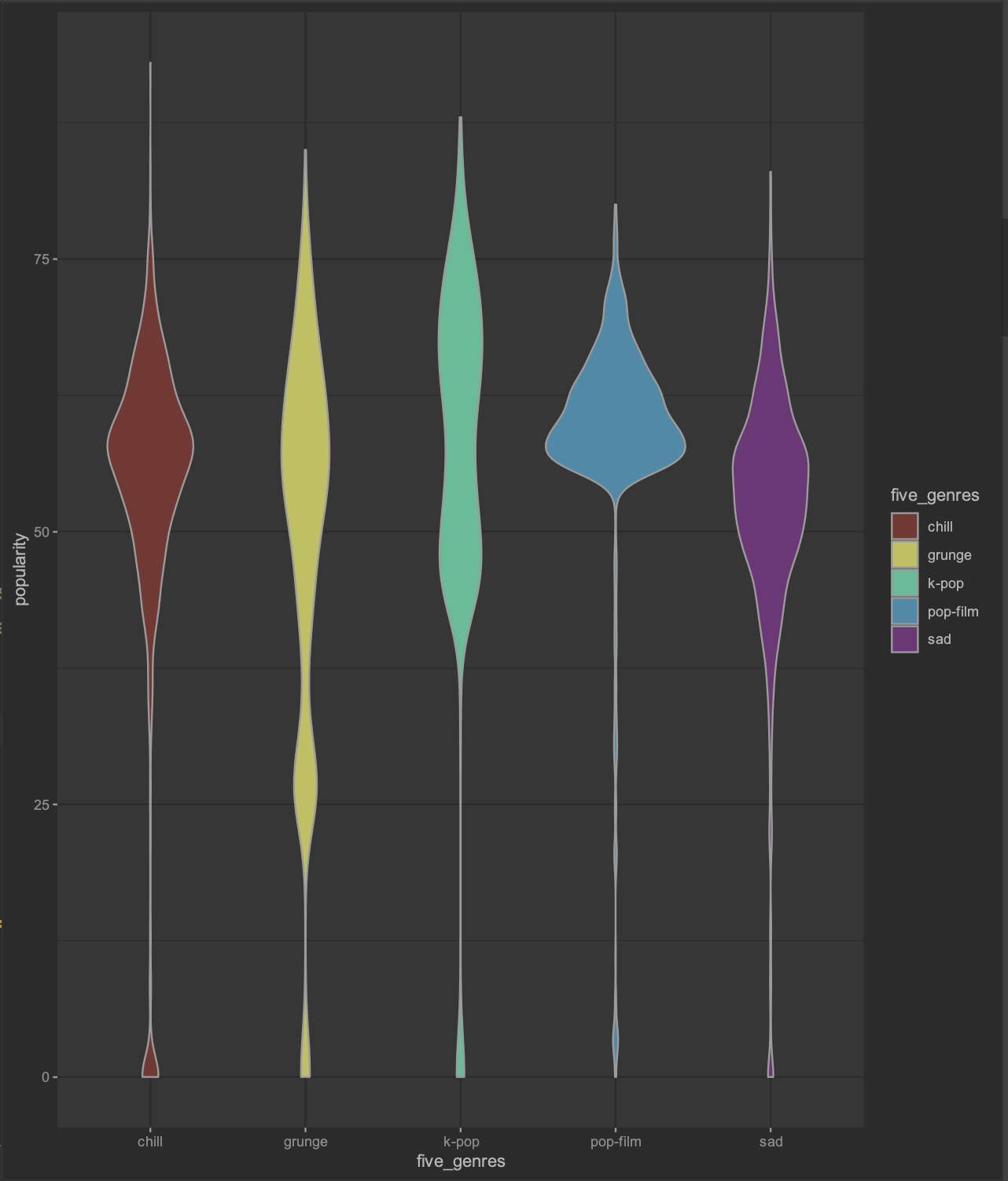
most\_popular\_albums виконує ті самі операції, що й most\_popular\_artists, але з назвами альбомів замість виконавців. Вибираються 20 найкращих альбомів і наносяться на lollipop діаграму.



most\_popular\_genres обчислює загальний бал популярності для кожного жанру треків і вибирає 5 найпопулярніших жанрів. Отриманий фрейм даних роздруковується, а для візуалізації даних за допомогою ggplot створюється гістограма.

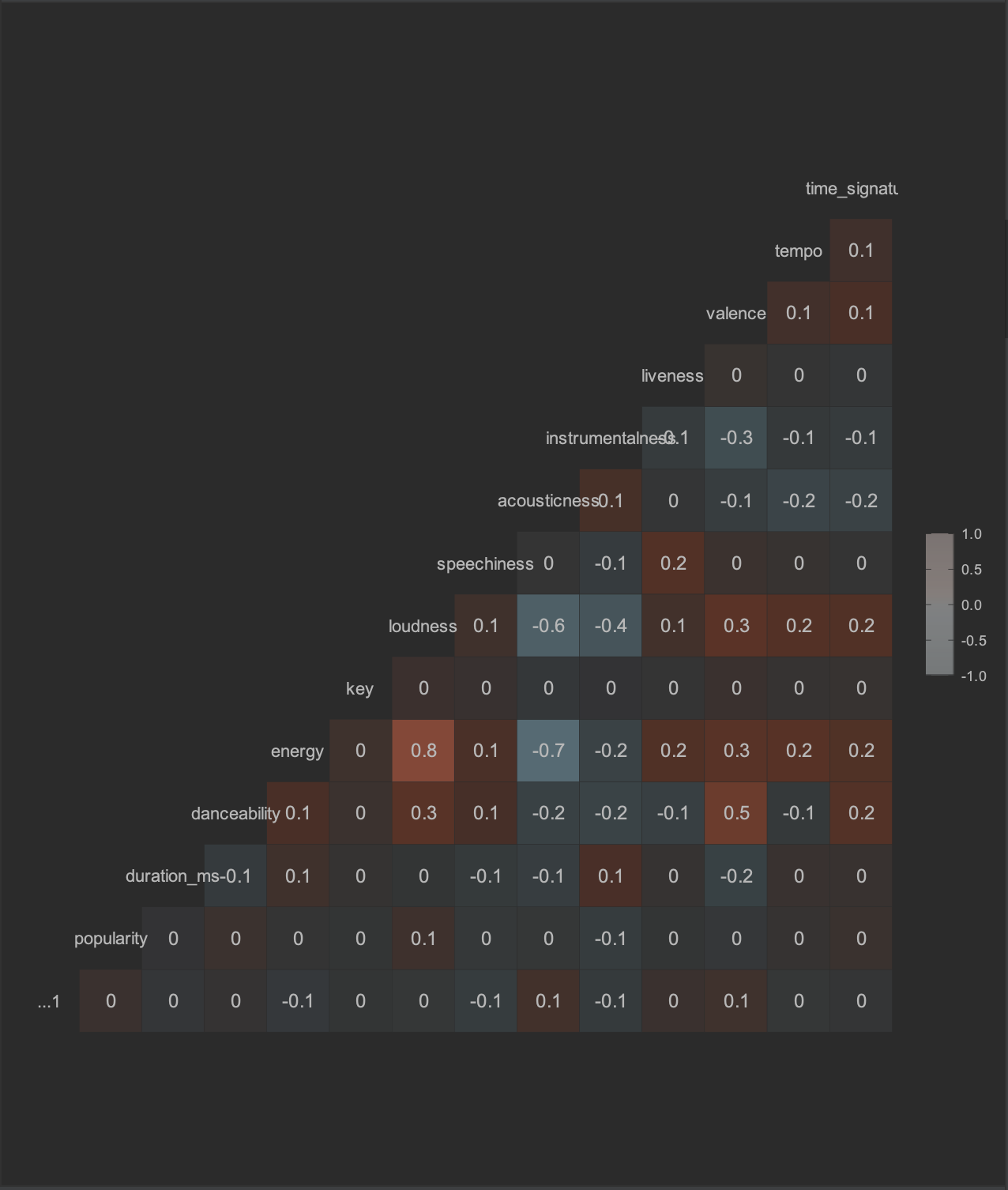


Далі ми підготовуємо дані для побудови violan plot, фільтруючи їх, щоб включити лише пісні п'яти конкретних жанрів 'pop-film', 'k-pop', 'chill', 'sad', and 'grunge'. Потім ми створюємо violan plot за допомогою ggplot для візуалізації розподілу змінної "популярність" за цими п'ятьма жанрами. Аргумент "fill" використовується для розфарбовування скрипок відповідно до жанру, який вони представляють.



**Графік кореляціїї числових неперервних змінних**

Створюємо кореляційний графік для числових змінних у наборі даних за допомогою функції ggcorr(). Цей графік відображає парні кореляції між кожною парою змінних, причому сила кореляції позначається кольором та інтенсивністю відповідного квадрата. Змінні, які мають високу кореляцію (позитивну або негативну), згруповані разом.



По результатам графіку було виявлено, що змінні "energy" і "loudness" мають високу позитивну кореляцію (0,8). Це означає, що зі збільшенням рівня енергії пісні зростає і її гучність.

Графік кореляції також показує, що енергія має негативну кореляцію з акустичністю (кореляція -0,7). Це означає, що зі збільшенням енергії треку його акустичність має тенденцію до зменшення, і навпаки. Ця інформація може бути корисною при аналізі та розумінні характеристик різних типів музики.

По цьому графіку можно робити певні виводи щодо поставлених питань:

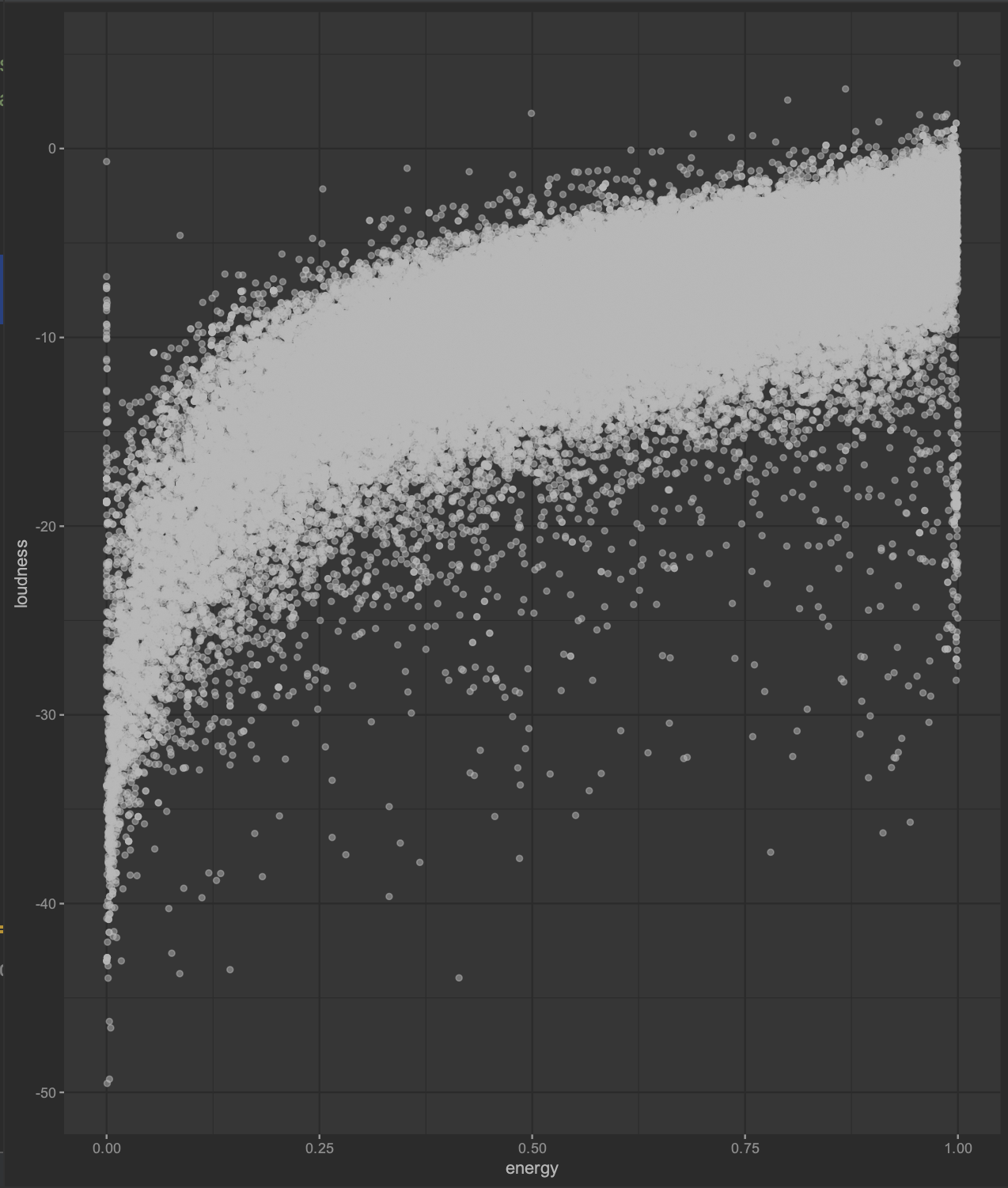
* Залежність популярності треку за валентністю(valence)
* Як тривалість треку впливає на його енергійність

Виходячи з результатів кореляційного графіка, можна зробити висновок, що між проаналізованими нами змінними не існує значущої кореляції. Зокрема, немає кореляції між популярністю треку та валентністю, довжиною треку та його енергічністю, ненормативною лексикою та темпом, що впливає на промовистість треку. Значення кореляції для цих показників дорівнювали 0, що свідчить про відсутність зв'язку.

Далі генерує діаграму розсіювання для візуалізації зв'язку між двома змінними, "енергією" та "гучністю".

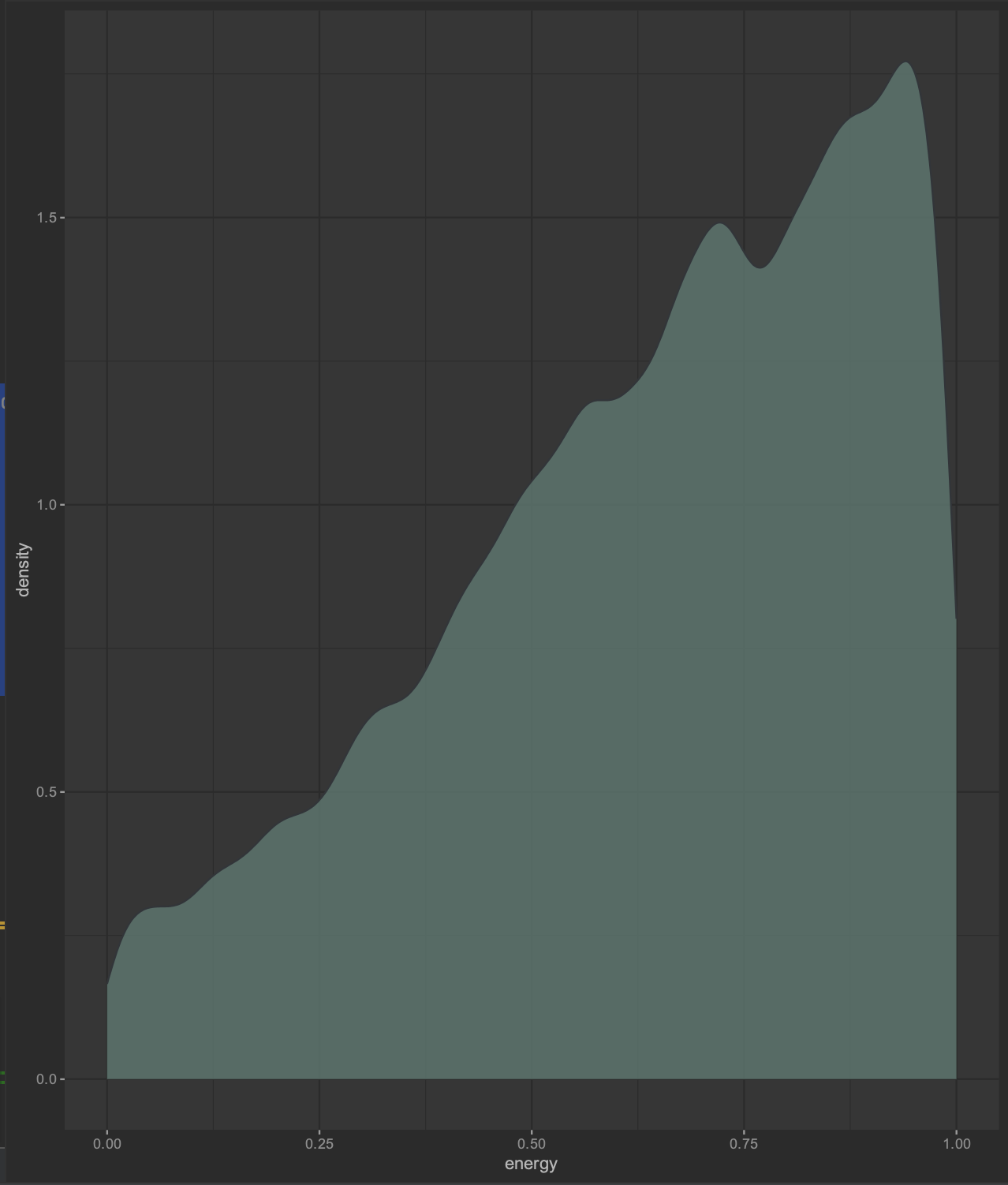
На графіку показано рівень енергії кожної доріжки на горизонтальній осі та рівень гучності на вертикальній осі.

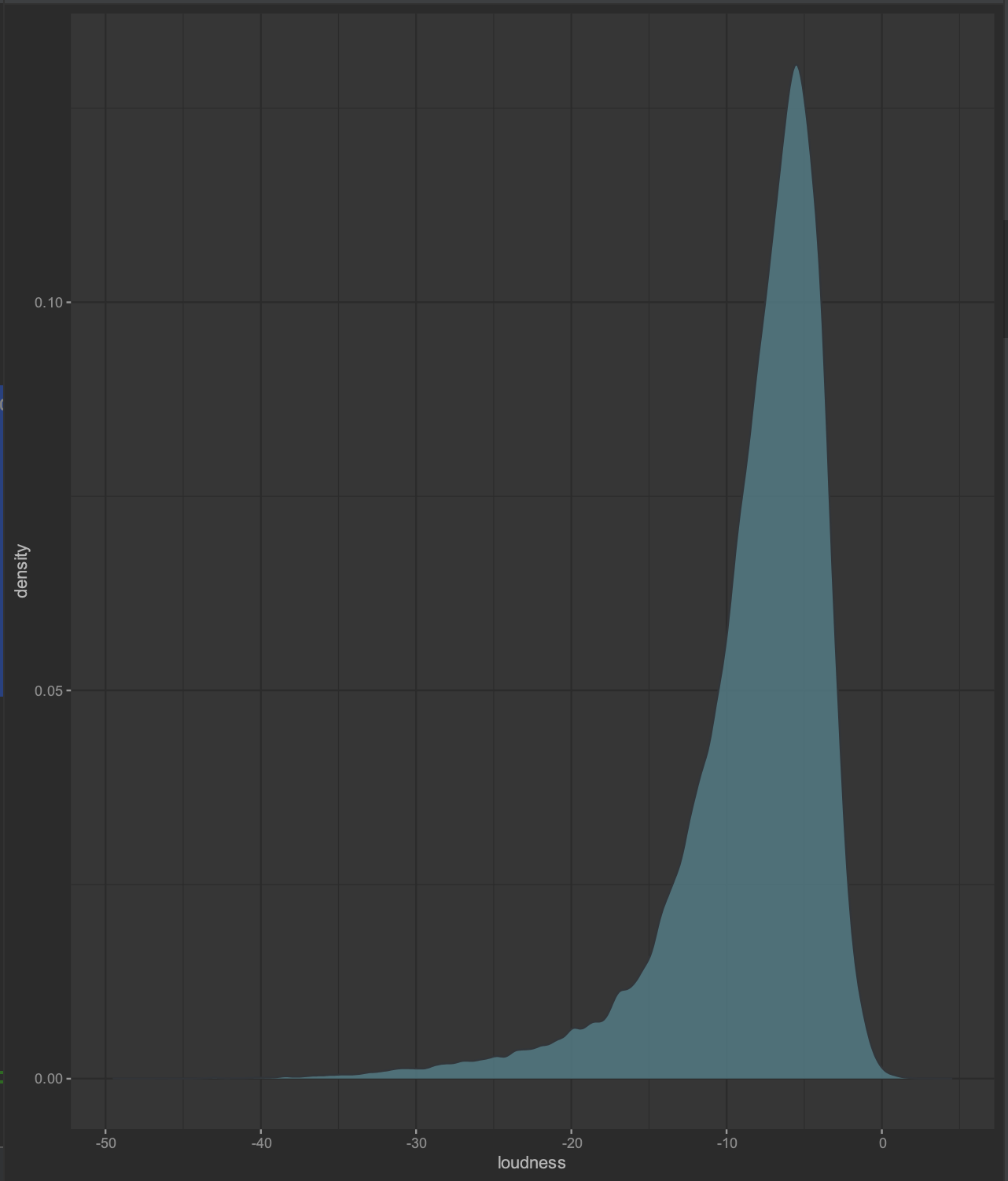
Вивчаючи графік, ми можемо візуально підтвердити існування кореляції між рівнями енергії та гучності треків.



**графіки щільності для змінних "енергія" та "гучність"**

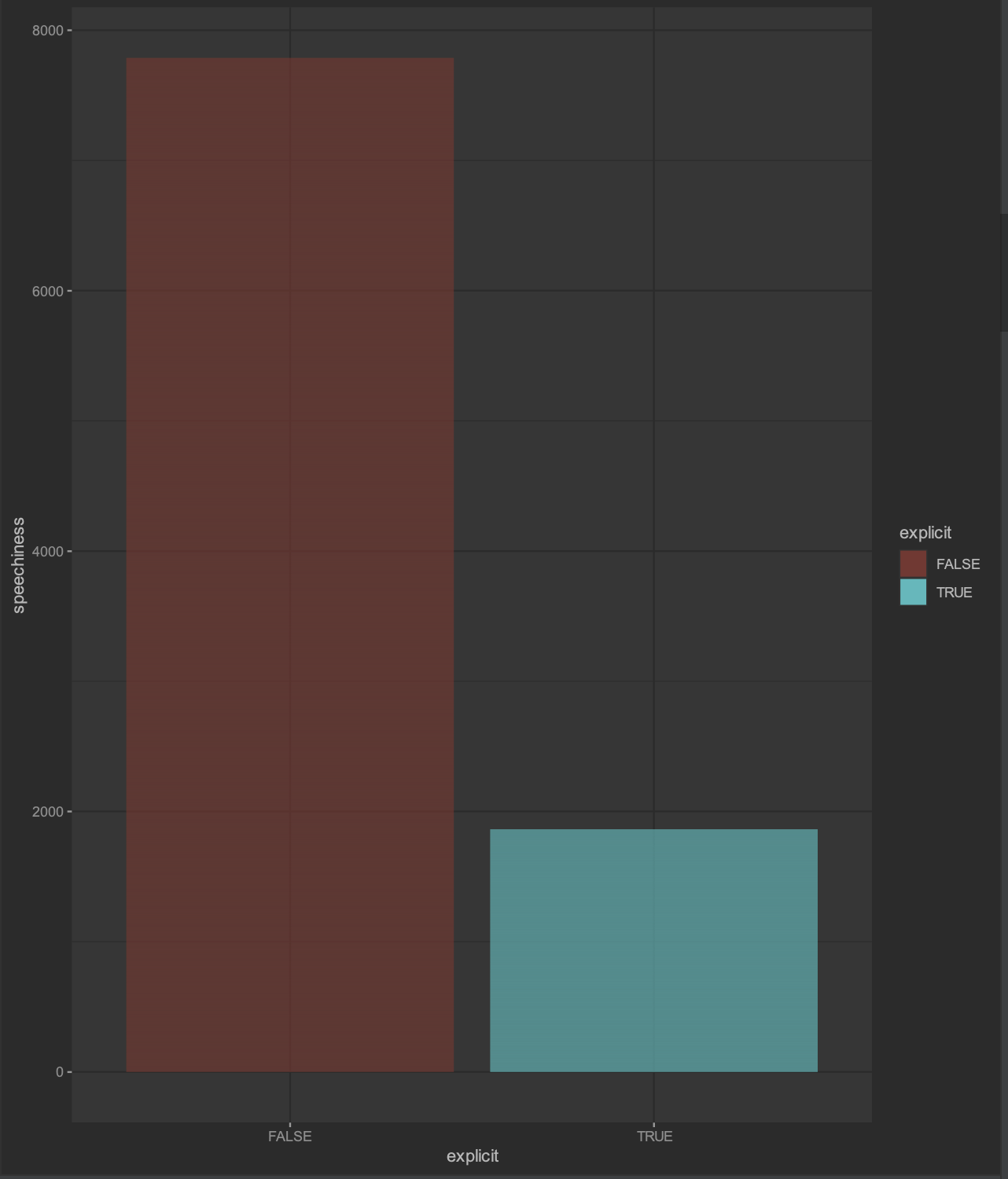
Побудувавши 2 графіки щільності для змінних "енергія" та "гучність" відповідно, ми також можемо спостерігати певну залежність між цими змінними. Графік щільності є представленням розподілу неперервної змінної і схожий на гістограму.





**Стовпчаста діаграма speechiness для кожної групи explicit**

Генеруємо гістограму, яка відображає мовність доріжок у кожній явній групі. Доріжки поділено на дві групи, експліцитні та не експліцитні, а мовленнєвість показано на осі y. На осі х відкладено експліцитні та не експліцитні групи, які позначено різними кольорами. Висота кожного стовпчика відображає мовленнєву активність доріжок у цій групі, причому вищі стовпчики вказують на вищу мовленнєву активність.



Цей графік дає нам змогу відповісти на поставлене питання:

* Як ненормативна лексика й темп впливає на мовність(speechiness) треку

На графіку можна побачити, що в категорії, де не має ненормативної лексики - мовність набагато більша.

Отже робимо висновок, що дійсно є висока залежність між цими змінними. Тобто, наші очікування дійсно підтвердились.

**Висновки**

1. Початкові питання, які були задані щодо набору даних, були цікавими та релевантними, а аналіз даних допоміг нам краще зрозуміти взаємозв'язки між різними змінними.
2. Ми використовували різні методи візуалізації, щоб дослідити дані та відповісти на початкові запитання. Зокрема, ми використовували діаграми розсіювання, кореляційні матриці, скрипкові діаграми та діаграми щільності для візуалізації взаємозв'язків між різними змінними.
3. Виходячи з результатів аналізу, можна зробити висновок, що між популярністю треку та валентністю, довжиною треку та енергійністю, не існує значущої кореляції. Значення кореляції для цих показників дорівнювали 0, що свідчить про відсутність зв'язку.

Ми також виявили, що існує висока залежність між ненормативною лексикою та мовністю треку. Зокрема, в категорії, де немає ненормативної лексики, мовність є набагато вищою. Це свідчить про те, що використання ненормативної лексики може впливати на мовність треку.

1. Однак ми виявили сильну негативну кореляцію між енергійністю та акустичністю, а також сильну позитивну кореляцію між енергійністю та гучністю. Ці результати свідчать про те, що високоенергетичні треки, як правило, є більш гучними і менш акустичними.

Загалом, наш аналіз музичних даних дозволив нам дослідити взаємозв'язки між різними змінними та відповісти на початкові запитання. Хоча ми не виявили значущих кореляцій між усіма досліджуваними змінними, ми виявили цікаві зв'язки між деякими змінними, які варті подальшого вивчення.