КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

імені ТАРАСА ШЕВЧЕНКА

Факультет комп'ютерних наук та кібернетики

Звіт

з дисципліни «Аналіз даних»

про виконання завдання «Попередній аналіз даних»

Виконав:

студент групи К-28

Миронов Данiiл Олександрович

Київ-2022

Зміст

tbd

**Опис вхідної інформації**

**Основнi данi**

Для аналiзу вибрано набiр даних «TikTok Trending Tracks».

Посилання на джерело даних: <https://www.kaggle.com/yamqwe/tiktok-trending-tracks>

Набір даних містить популярні треки, представлені в сервiсi TikTok. Цей набір даних містить близько 7000 треків разом із даними про них, оцінкою характеристик, таких як гучнiсть, енергiчнiсть, iншi данi та технічнe інформацію.

Набiр був створений командою Team Dan. Посилання на автора набору: <https://github.com/romeoben/DSC7-Sprint2-TeamDan>

**Опис змiнних**

У наборі даних присутні 23 змінні:

|  |  |
| --- | --- |
| **Назва змінної** | **Опис** |
| track\_id | ID треку в системі TikTok |
| track\_name | Назва треку |
| artist\_id | ID виконавця в системі TikTok |
| artist\_name | Ім’я виконавця |
| album\_id | ID альбому в системі TikTok |
| duration | Тривалість треку в мілісекундах |
| release\_date | Дата виходу треку |
| popularity | Популярність треку – ціле число на шкалі від 0 до 100 включно |
| danceability | Танцювальність треку – оцінка від 0 до 1 |
| energy | Енергічність треку – оцінка від 0 до 1 |
| key | Ключ виконання треку |
| loudness | Гучність треку |
| mode | Режим треку |
| speechiness | Мовленнєвість треку – оцінка від 0 до 1 |
| acousticness | Акустичність треку – оцінка від 0 до 1 |
| instrumentalness | Інструментальність треку – оцінка від 0 до 1 |
| liveness | Живучість треку – оцінка від 0 до 1 |
| valence | Музична позитивність, яку передає трек – оцінка від 0 до 1 |
| tempo | Темп треку – удари в хвилину |
| playlist\_id | ID плейлисту в системі TikTok |
| playlist\_name | Назва плейлисту |
| duration\_mins | Тривалість треку в хвилинах |
| genre | Жанр треку |

**Постановка задачі**

Загальна мета задачі - провести попередній аналіз обраного набору даних.

Детальне формулювання:

Для кожної зі змінних *duration*, *popularity*, *energy*, *tempo*:

* дати її класифікацію,
* графічно представити (емпіричну функцію щільності)/(полігон частот),
* побудувати зображення "скринька з вусами",
* підрахувати вибіркові значення: мінімального та максимального спостережень вибірки, медіани, квартилів, децилів,
* підрахувати вибіркові значення характеристик положення центру значень,
* підрахувати вибіркові значення характеристик розсіювання значень,
* провести аналіз скошеності та гостроверхості розподілу,
* провести аналіз отриманих результатів передньої обробки даних та сформулювати відповідні висновки

**Передмова до аналізу**

Цей звіт, виконувані файли з розширенням .R, результати виконання та інші лабораторні роботи з поточним набором даних доступні за посиланням: <https://github.com/leirimnad/data-analysis-labs>

Перед виконанням коду з наступних розділів, матимемо на увазі, що перед кожним з фрагментів виконується наступний код:

options(scipen=5)

songs <- read.csv(file = 'tiktok.csv')

Фрагмент коду options(scipen=5) використовується для того, щоб числові дані виводилися не в експоненціальному записі.

Фрагмент songs <- read.csv(file = 'tiktok.csv') обробляє .csv-файл з даними та зберігає їх у змінну songs.

**Аналіз змінної *duration***

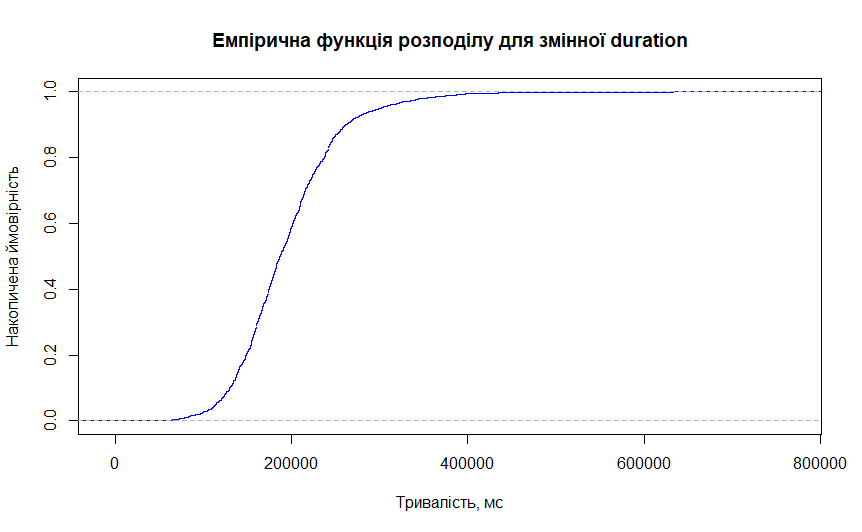
**Опис змінної:** тривалість треку в мілісекундах.

**Класифікація:** змінна скалярна, кількісна.

**Зауваження:** домовимося, що перед виконанням коду з цього розділу було виконано наступний код:

durations <- songs$duration

**Графічне представлення змінної**



Код для виводу емпіричної функції розподілу:

plot(ecdf(durations),

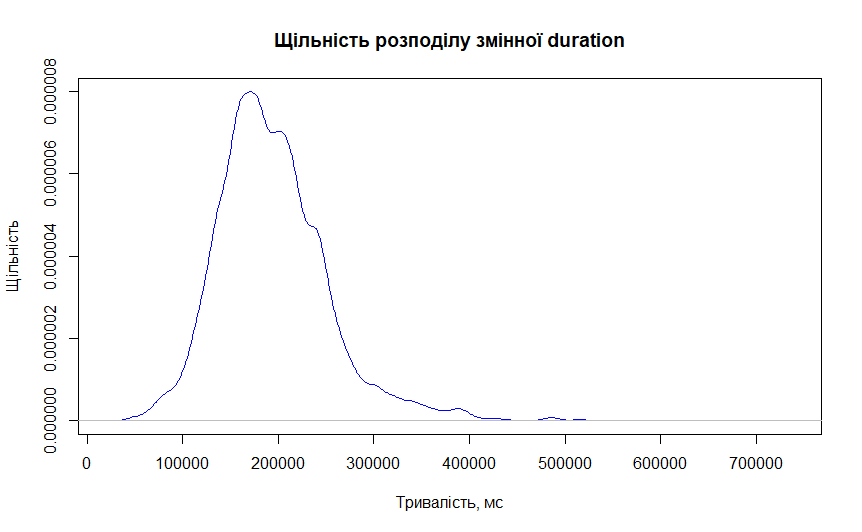
main="Емпірична функція розподілу для змінної duration",

xlab="Тривалість, мс",

ylab="Накопичена ймовірність",

col = "blue"

)



Код для виводу щільності розподілу:

plot(density(durations),

main="Щільність розподілу змінної duration",

xlab="Тривалість, мс",

ylab="Щільність",

col = "blue"

)

**Зображення "скринька з вусами"**

Код для виводу зображення "скринька з вусами":

boxplot(

durations,

main = "",

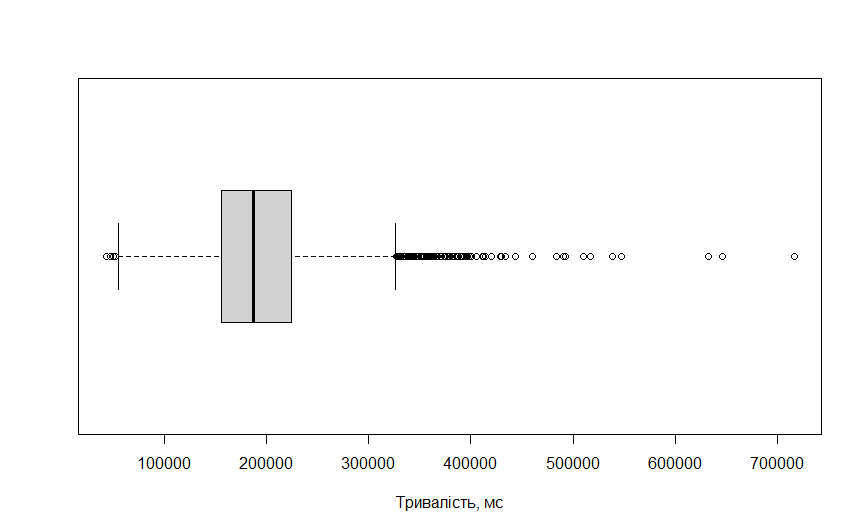
xlab = "Тривалість, мс",

ylab = "",

horizontal = TRUE,

notch = FALSE

)

Результат:

**Вибіркові значення: мінімальне та максимальне спостереження вибірки, медіана, квартилі, децилі**

|  |  |  |
| --- | --- | --- |
| Характеристика | Значення | Код, скріншот виконання коду |
| Мінімальне спостереження вибірки | 43426 | summary(durations) |
| Максимальне спостереження вибірки | 716206 |
| Медіана | 186980 |
| Квартилі | Нижній: 155866  Верхній: 224284 |
| Децилі | 1. 129543.5 2. 148345.0 3. 161420.5 4. 174548.0 5. 186980.0 6. 201398.0 7. 215382.0 8. 237106.0 9. 261866.0 | quantile(durations, probs = seq(.1, .9, by = .1)) |

**Вибіркові значення характеристик положення центру значень**

|  |  |  |
| --- | --- | --- |
| Характеристика | Значення | Код, скріншот виконання коду |
| Вибіркове середнє арифметичне | 194428.739 | print(paste("Вибіркове середнє арифметичне: ", mean(durations))) |
| Середнє геометричне | 186005.387 | print(paste("Середнє геометричне: ", exp(mean(log(durations))))) |
| Середнє гармонічне | 177539.247 | print(paste("Середнє гармонічне: ", 1/mean(1/durations))) |
| Мода | 160000 | Mode <- function(x) {  ux <- unique(x)  ux[which.max(tabulate(match(x, ux)))]  }  print(paste("Мода: ", Mode(durations))) |
| Медіана | 186980 | print(paste("Медіана: ", median(durations))) |

**Вибіркові значення характеристик розсіювання значень**

|  |  |  |
| --- | --- | --- |
| Характеристика | Значення | Код, скріншот виконання коду |
| Дисперсія | 3505657540.7 | print(paste("Дисперсія:", var(durations))) |
| Стандартне відхилення | 59208.593 | print(paste("Стандартне відхилення:", sd(durations))) |
| Коефіцієнт варіації | 30.45% | print(paste("Коефіцієнт варіації:", sd(durations) / mean(durations) \* 100, "%")) |
| Ймовірнісне відхилення | 190075 | print(paste("Ймовірнісне відхилення: ", mean(quantile(durations, probs=c(.25, .75))))) |
| Розмах вибірки | 672780 | print(paste("Розмах вибірки: ", max(durations)-min(durations))) |
| Інтервал концентрації розподілу | (16802.959, 372054.52) | print(paste("Інтервал концентрації розподілу: (",  mean(durations)-3\*sd(durations),  ",",  mean(durations)+3\*sd(durations),  ")"  )) |

**Аналіз скошеності та гостроверхості розподілу**

Для аналізу скошеності та гостроверхості використовуємо бібліотеку moments:

install.packages("moments")

library(moments)

Код для підрахунку *коефіцієнту асиметрії*:

> print(paste("Коефіцієнт асиметрії:", skewness(durations)))

[1] "Коефіцієнт асиметрії: 1.22250165941111"

Коефіцієнт асиметрії більший за нуль, отже розподіл скошений ліворуч.

Код для підрахунку *коефіцієнту ексцесу*:

> print(paste("Коефіцієнт ексцесу:", kurtosis(durations)))

[1] "Коефіцієнт ексцесу: 7.22533147640749"

Коефіцієнт ексцесу більший за нуль, отже розподіл більш гостроверхий ніж нормальний с відповідними параметрами.

Для більш явного порівняння розподілу значень змінної duration з відповідним нормальним розподілом, виконаємо наступний код:

plot(density(durations),

col="blue",

main="Порівняння щільності duration з відповідним нормальним розподілом",

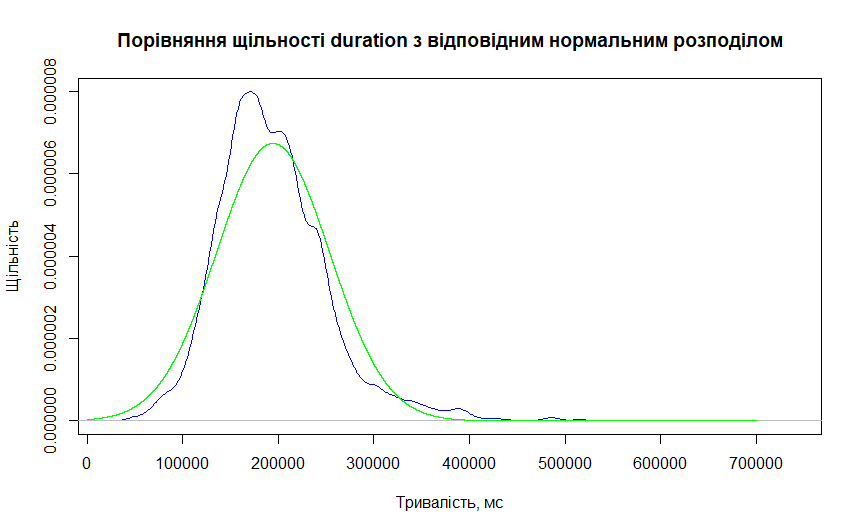
xlab="Тривалість, мс",

ylab="Щільність",

)

lines(dnorm(x=seq(0, 700000), mean=mean(durations), sd=sd(durations)),col="green")

Результат:



Зеленим кольором зображено щільність нормального розподілу з відповідними параметрами.

**Висновки попереднього аналізу змінної duration**

На основі підрахованих в цьому розділі характеристик тривалості треку можемо зробити наступні висновки:

* В середньому популярні треки мають тривалість близько 3 хвилин
* Найчастіше зустрічаються треки довжиною рівно 160 секунд
* Довжина треків має досить великий розмах, зустрічаються треки від 43 секунд до 12 хвилин
* Однак, треки довжиною більше ніж п’ять з половиною хвилин скоріш є викидами

**Аналіз змінної *popularity***

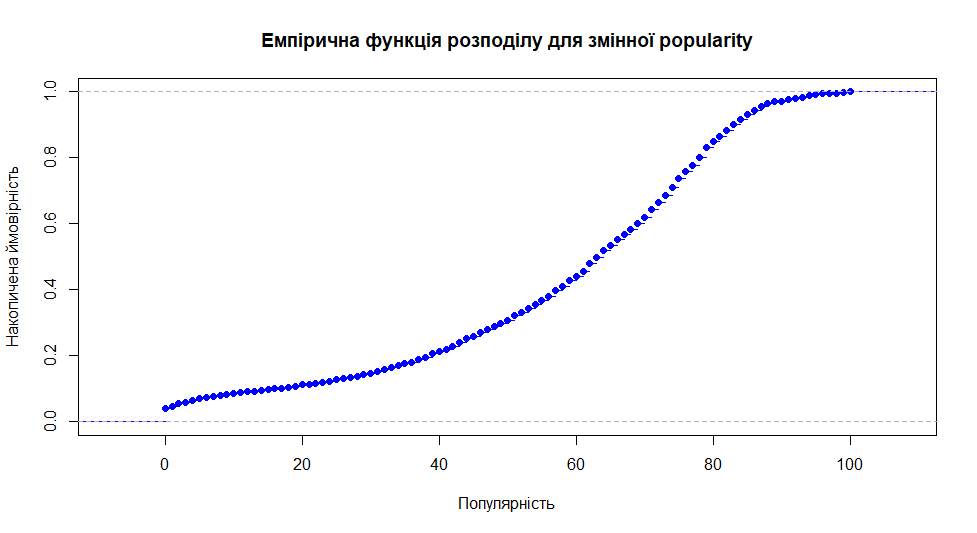
**Опис змінної:** популярність треку – ціле число на шкалі від 0 до 100 включно.

**Класифікація:** змінна скалярна, кількісна.

**Зауваження:** домовимося, що перед виконанням коду з цього розділу було виконано наступний код:

popularity <- songs$popularity

**Графічне представлення змінної**



Код для виводу емпіричної функції розподілу:

plot(ecdf(popularity),

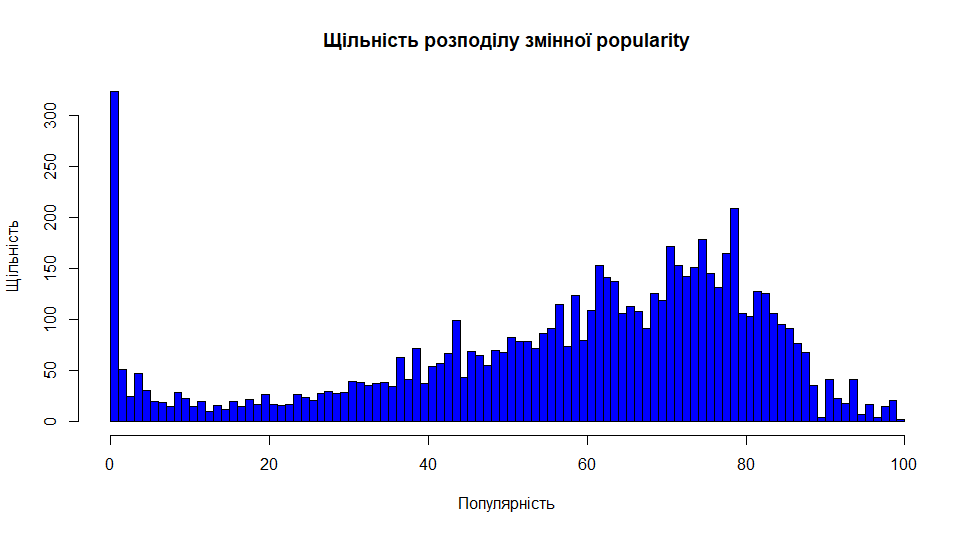
main="Емпірична функція розподілу для змінної popularity",

xlab="Популярність",

ylab="Накопичена ймовірність",

col = "blue"

)



Код для виводу щільності розподілу у вигляді гістограми:

plot(hist(popularity, breaks=100),

main="Щільність розподілу змінної popularity",

xlab="Популярність",

ylab="Щільність",

col = "blue"

)

**Зображення "скринька з вусами"**

Код для виводу зображення "скринька з вусами":

boxplot(

popularity,

main = "",

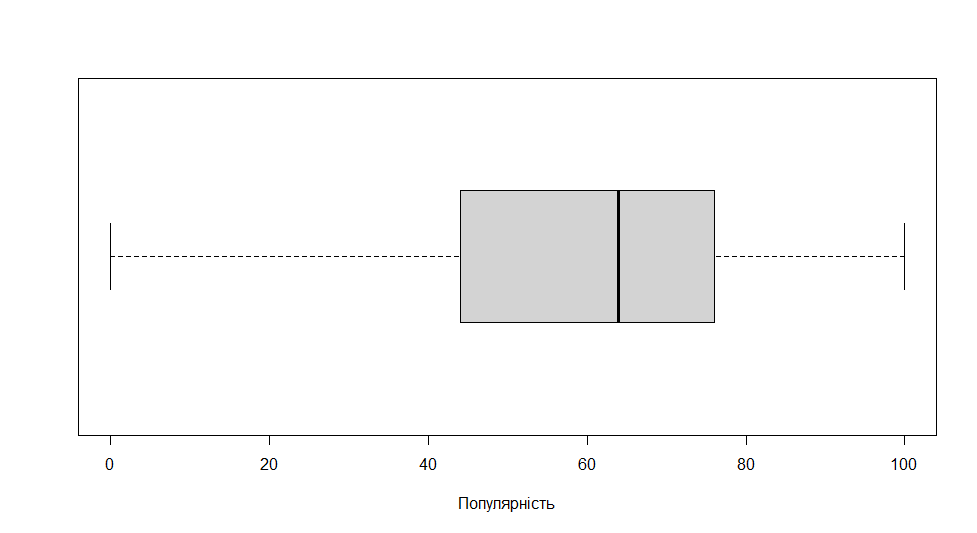
xlab = "Популярність",

ylab = "",

horizontal = TRUE,

notch = FALSE

)

Результат:

**Вибіркові значення: мінімальне та максимальне спостереження вибірки, медіана, квартилі, децилі**

|  |  |  |
| --- | --- | --- |
| Характеристика | Значення | Код, скріншот виконання коду |
| Мінімальне спостереження вибірки | 0 | summary(popularity) |
| Максимальне спостереження вибірки | 100 |
| Медіана | 64 |
| Квартилі | Нижній: 44  Верхній: 76 |
| Децилі | 1. 16 2. 39 3. 50 4. 58 5. 64 6. 69 7. 74 8. 78 9. 83 | quantile(popularity, probs = seq(.1, .9, by = .1)) |

**Вибіркові значення характеристик положення центру значень**

|  |  |  |
| --- | --- | --- |
| Характеристика | Значення | Код, скріншот виконання коду |
| Вибіркове середнє арифметичне | 57.65 | print(paste("Вибіркове середнє арифметичне: ", mean(popularity))) |
| Середнє геометричне | 0 | print(paste("Середнє геометричне: ", exp(mean(log(popularity))))) |
| Середнє гармонічне | 0 | print(paste("Середнє гармонічне: ", 1/mean(1/popularity))) |
| Мода | 0 | Mode <- function(x) {  ux <- unique(x)  ux[which.max(tabulate(match(x, ux)))]  }  print(paste("Мода: ", Mode(popularity))) |
| Медіана | 64 | print(paste("Медіана: ", median(popularity))) |

**Вибіркові значення характеристик розсіювання значень**

|  |  |  |
| --- | --- | --- |
| Характеристика | Значення | Код, скріншот виконання коду |
| Дисперсія | 605.97 | print(paste("Дисперсія:", var(popularity))) |
| Стандартне відхилення | 24.616 | print(paste("Стандартне відхилення:", sd(popularity))) |
| Коефіцієнт варіації | 42.7% | print(paste("Коефіцієнт варіації:", sd(popularity) / mean(popularity) \* 100, "%")) |
| Ймовірнісне відхилення | 60 | print(paste("Ймовірнісне відхилення: ", mean(quantile(popularity, probs=c(.25, .75))))) |
| Розмах вибірки | 100 | print(paste("Розмах вибірки: ", max(popularity)-min(popularity))) |
| Інтервал концентрації розподілу | (-16.2, 131.5) | print(paste("Інтервал концентрації розподілу: (",  mean(popularity)-3\*sd(popularity),  ",",  mean(popularity)+3\*sd(popularity),  ")"  )) |

**Аналіз скошеності та гостроверхості розподілу**

Для аналізу скошеності та гостроверхості використовуємо бібліотеку moments:

install.packages("moments")

library(moments)

Код для підрахунку *коефіцієнту асиметрії*:

> print(paste("Коефіцієнт асиметрії:", skewness(popularity)))

[1] "Коефіцієнт асиметрії: -0.892121349433045"

Коефіцієнт асиметрії менший за нуль, отже розподіл скошений праворуч.

Код для підрахунку *коефіцієнту ексцесу*:

> print(paste("Коефіцієнт ексцесу:", kurtosis(popularity)))

[1] "Коефіцієнт ексцесу: 2.99387403923279"

Коефіцієнт ексцесу більший за нуль, отже розподіл більш гостроверхий ніж нормальний с відповідними параметрами.

Для більш явного порівняння розподілу значень змінної popularity з відповідним нормальним розподілом, виконаємо наступний код:

plot(density(popularity),

col="blue",

main="Порівняння щільності popularity з відповідним нормальним розподілом",

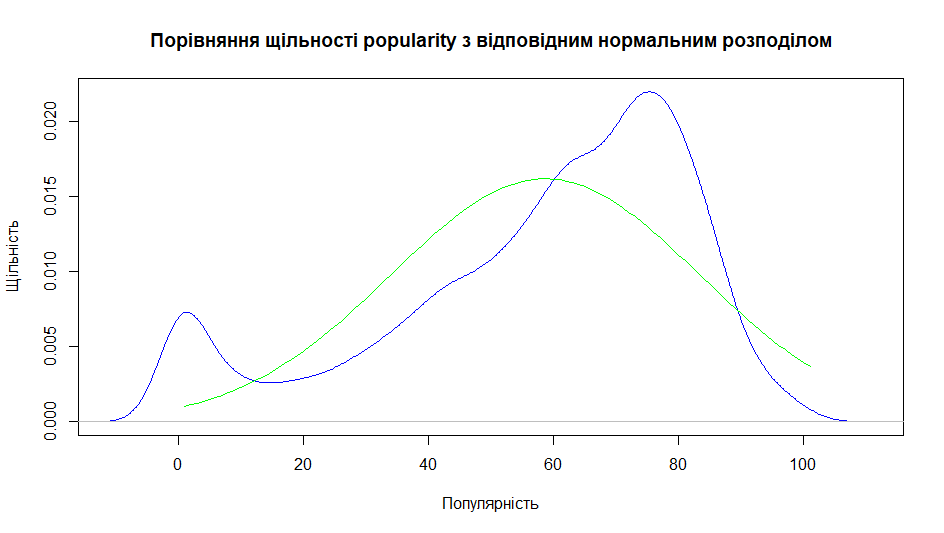
xlab="Популярність",

ylab="Щільність",

)

lines(dnorm(x=seq(0, 100), mean=mean(popularity), sd=sd(popularity)),col="green")

Результат:



Зеленим кольором зображено щільність нормального розподілу з відповідними параметрами.

**Висновки попереднього аналізу змінної popularity**

На основі підрахованих в цьому розділі характеристик популярності треку можемо зробити наступні висновки:

* Досить багато значень популярності дорівнюють нулю, при тому що надалі кількість треків з малою популярністю різко знижується
* Середнє значення популярності більше ніж 50