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*:

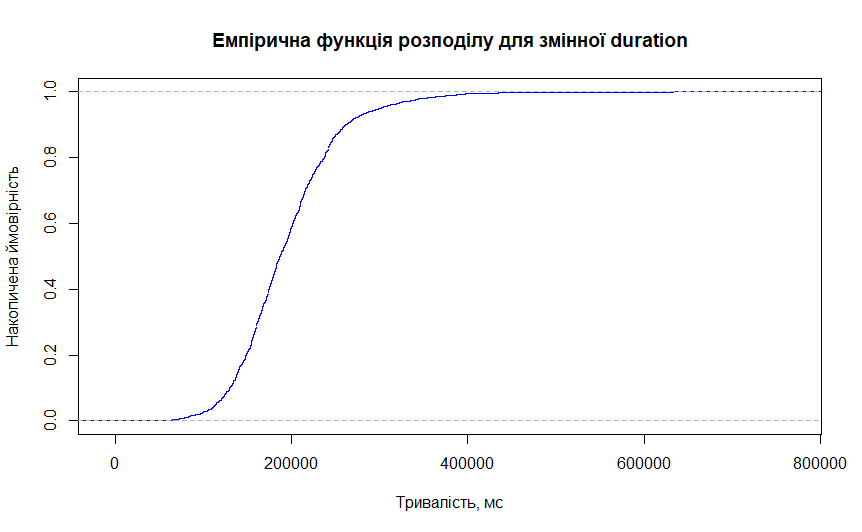
* дати її класифікацію,
* графічно представити (емпіричну функцію щільності)/(полігон частот),
* побудувати зображення "скринька з вусами",
* підрахувати вибіркові значення: мінімального та максимального спостережень вибірки, медіани, квартилів, децилів,
* підрахувати вибіркові значення характеристик 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

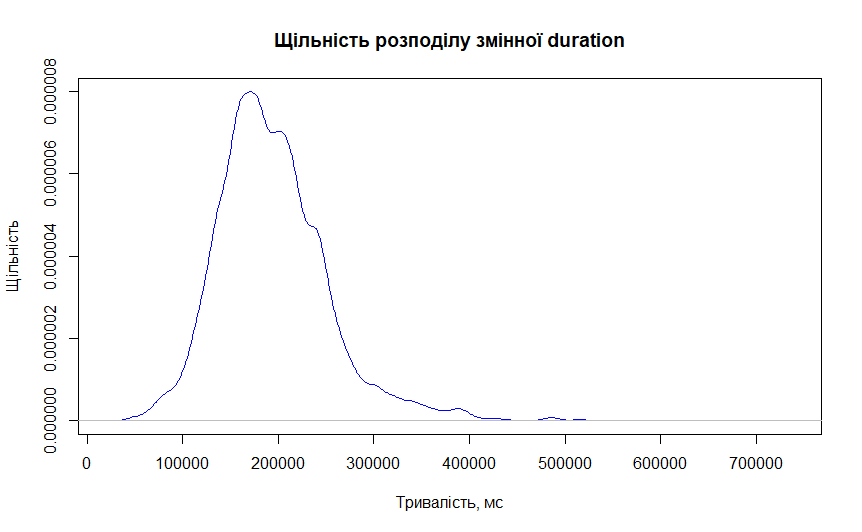
**Аналіз змінної *duration***

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

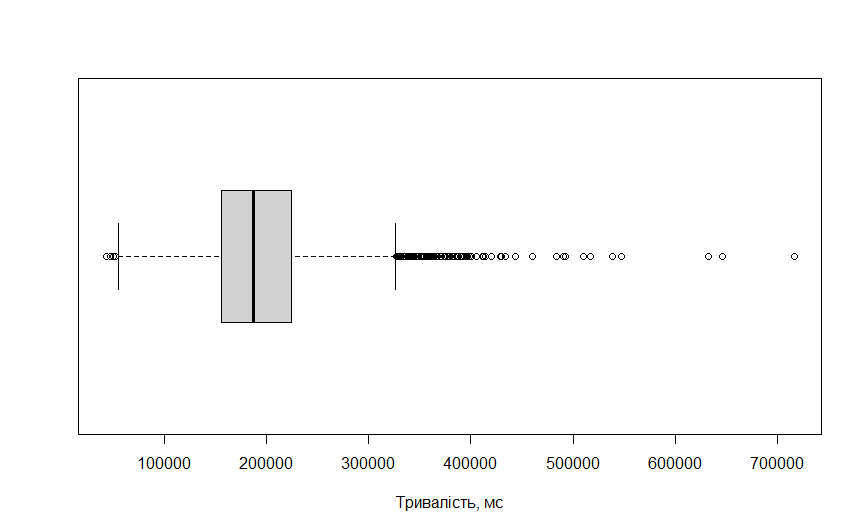
**Класифікація:** змінна скалярна, кількісна.

**Графічне представлення змінної**





**Зображення "скринька з вусами"**



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

> summary(durations)

Min. 1st Qu. Median Mean 3rd Qu. Max.

43426 155866 186980 194429 224284 716206

> quantile(durations, probs = seq(.1, .9, by = .1))

10% 20% 30% 40% 50% 60% 70% 80% 90%

129543.5 148345.0 161420.5 174548.0 186980.0 201398.0 215382.0 237106.0 261866.0

**Вибіркові значення характеристик положення центру значень**

> print(paste("Мода: ", Mode(durations)))

[1] "Мода: 160000"

> print(paste("Медіана: ", median(durations)))

[1] "Медіана: 186980"

> print(paste("Вибіркове середнє арифметичне: ", mean(durations)))

[1] "Вибіркове середнє арифметичне: 194428.739252891"

> print(paste("Середнє геометричне: ", exp(mean(log(durations)))))

[1] "Середнє геометричне: 186005.387491742"

> print(paste("Середнє гармонічне: ", 1/mean(1/durations)))

[1] "Середнє гармонічне: 177539.247105408"

>

> Mode <- function(x) {

+ ux <- unique(x)

+ ux[which.max(tabulate(match(x, ux)))]

+ }

>

> print(paste("Мода: ", Mode(durations)))

[1] "Мода: 160000"

> print(paste("Медіана: ", median(durations)))

[1] "Медіана: 186980"

**Вибіркові значення характеристик розсіювання значень**

> print(paste("Дисперсія:", var(durations)))

[1] "Дисперсія: 3505657540.66425"

> print(paste("Стандартне відхилення:", sd(durations)))

[1] "Стандартне відхилення: 59208.5934697342"

> print(paste("Коефіцієнт варіації:", sd(durations) / mean(durations) \* 100, "%"))

[1] "Коефіцієнт варіації: 30.452593426902 %"

> print(paste("Ймовірнісне відхилення: ", mean(quantile(durations, probs=c(.25, .75)))))

[1] "Ймовірнісне відхилення: 190075"

> print(paste("Розмах вибірки: ", max(durations)-min(durations)))

[1] "Розмах вибірки: 672780"

> print(paste("Інтервал концентрації розподілу: (",

+ mean(durations)-3\*sd(durations),

+ ",",

+ mean(durations)+3\*sd(durations),

+ ")"

+ ))

[1] "Інтервал концентрації розподілу: ( 16802.9588436881 , 372054.519662093 )"

**Аналіз скошеності та гостроверхості розподілу**

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

[1] "Коефіцієнт асиметрії: 1.22250165941111"

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

[1] "Коефіцієнт ексцесу: 7.22533147640749"

