КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

імені ТАРАСА ШЕВЧЕНКА

Факультет комп'ютерних наук та кібернетики

Звіт

з дисципліни «Аналіз даних»

про виконання завдання «Кореляційний аналіз даних»

Виконав:

студент групи К-28

Миронов Данiiл Олександрович

Київ-2022

Зміст

[Зміст 2](#_Toc102312168)

[Опис вхідної інформації 3](#_Toc102312169)

[Постановка задачі 5](#_Toc102312170)

[Передмова до аналізу 6](#_Toc102312171)

[Вибір характеристик статистичного зв’язку 7](#_Toc102312172)

[Тест на нормальність розподілу 7](#_Toc102312173)

[Вибір характеристик статистичного зв’язку 7](#_Toc102312174)

[Аналіз істотності парних статистичних зв'язків 8](#_Toc102312175)

[Функція підрахунку індексу кореляції та перевірки його на істотність 8](#_Toc102312176)

[Функція підрахунку максимального рівня значущості при якому відповідний парний статистичний зв'язок не є значимим 8](#_Toc102312177)

[Підрахунок індексу кореляції для пар досліджуваних змінних 9](#_Toc102312178)

[Підрахунок максимального рівня значущості при якому парний статистичний зв'язок не є значимим 10](#_Toc102312179)

[Перевірка правильності підрахунку максимального рівня значущості 10](#_Toc102312180)

[Впорядкована послідовність пар змінних у порядку спадання істотності статистичного зв'язку між ними 11](#_Toc102312181)

[Аналіз істотності множинних статистичних зв'язків 12](#_Toc102312182)

[Функція підрахунку індексу кореляції та перевірки його на істотність 12](#_Toc102312183)

[Функція підрахунку максимального рівня значущості при якому відповідний множинний статистичний зв'язок не є значимим 12](#_Toc102312184)

[Підрахунок індексу кореляції для досліджуваних змінних 13](#_Toc102312185)

[Підрахунок максимального рівня значущості при якому множинний статистичний зв'язок не є значимим 14](#_Toc102312186)

[Перевірка правильності підрахунку максимального рівня значущості 14](#_Toc102312187)

[Впорядкована послідовність пар змінних у порядку спадання істотності множинного статистичного зв'язку їх з множиною усіх інших скалярних змінних 15](#_Toc102312188)

[Підсумок по лабораторній роботі 17](#_Toc102312189)

[Список використаних джерел 18](#_Toc102312190)

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

Провести кореляційний аналіз змінних ***duration***, ***popularity***, ***tempo***:

* на основі результатів попередньої обробки обраного набору даних визначитися, які характеристики статистичного зв'язку потрібно використати при подальшому їх кореляційному аналізі,
* провести аналіз істотності парних статистичних зв'язків для усіх пар скалярних змінних, навівши для них:
  + вибіркове значення відповідної парної характеристики статистичного зв'язку,
  + максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим,
  + впорядковану послідовність усіх пар скалярних змінних у порядку спадання істотності статистичного зв'язку між ними,
  + сформулювати висновки по кореляційному аналізу парних статистичних зв'язків для обраного набору скалярних змінних,
* провести аналіз істотності множинних статистичних зв'язків між кожною обраною в якості залежної скалярною змінною та множиною усіх інших скалярних змінних (які виступають у ролі незалежних змінних), навівши для них:
  + вибіркове значення відповідної множинної характеристики статистичного зв'язку,
  + максимальний рівень значущості при якому відповідний множинний статистичний зв'язок не є значимим,
  + впорядковану послідовність усіх скалярних змінних у порядку спадання істотності множинного статистичного зв'язку їх з множиною усіх інших
  + скалярних змінних,
  + сформулювати висновки по кореляційному аналізу множинних статистичних зв'язків для обраного набору скалярних змінних.

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.

Вибір характеристик статистичного зв’язку

Тест на нормальність розподілу

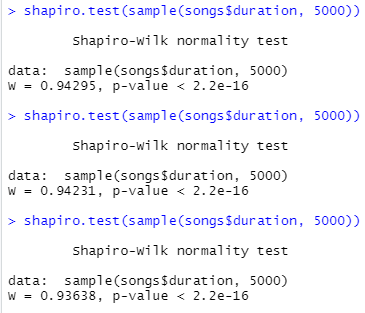
Для вибору доцільної характеристики проведемо дослідження на нормальність змінних. Для цього скористаємось тестом Шапіро-Вілка на нормальність, вбудованим в R. Визначимо нормальність для випадкової вибірки з 5000 значень для кожної зі змінних:

shapiro.test(sample(songs$duration, 5000))

shapiro.test(sample(songs$duration, 5000))

shapiro.test(sample(songs$duration, 5000))

Результат виконання:



Для кожної зі змінних значення p-value значно менше 0.05, отже значення змінних **не** є нормально розподіленими.

Вибір характеристик статистичного зв’язку

Зважаючи на те, що змінні не є нормально розподіленими, є кількісними, для дослідження парного статистичного зв’язку оберемо універсальну характеристику – **парний індекс кореляції**. Для дослідження множинного зв’язку також скористаємося **індексом кореляції**.

Аналіз істотності парних статистичних зв'язків

Функція підрахунку індексу кореляції та перевірки його на істотність

Для зручного аналізу пар змінних, створимо єдину функцію, яка буде вираховувати індекс кореляції, та перевіряти індекс кореляції на значимість.

Код функції наступний:

|  |
| --- |
| check\_correlation <- function(x, y, alpha=0.05){  n <- length(x)  p <- 2  indcor <- summary(lm(x ~ y))$r.squared  print(paste("Індекс кореляції:", indcor))    if(indcor == 0){  print("Зв'язку між змінними немає.")  } else if (indcor == 1){  print("Зв'язок між змінними функціональний.")  } else {    # Перевірка гіпотези I=0  f <- (indcor/(1-indcor))\*((n-p)/(p-1))  fa <- qf(1-alpha, p-1, n-p)    if(f < fa){  print("Зв'язок між змінними не є істотним.")  print("Прийнято гіпотезу про незначимість індексу кореляції:")  print(paste(f, "<", fa))  } else {  print("Зв'язок між змінними є істотним.")  print("Гіпотезу про незначимість індексу кореляції не прийнято:")  print(paste(f, ">=", fa))  }    }  return(indcor)  } |

Функція підрахунку максимального рівня значущості при якому відповідний парний статистичний зв'язок не є значимим

Створимо функцію, що повертатиме максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим. Скористаємось функцією оптимізації з бібліотеки *nloptr*, яка повертає максимальний рівень значущості, при якому виконується нерівність .

|  |
| --- |
| install.packages("nloptr")  library("nloptr")  max\_alpha <- function(x, y){  n <- length(x)  p <- 2  indcor <- summary(lm(x ~ y))$r.squared  f <- (indcor/(1-indcor))\*((n-p)/(p-1))    eval\_fa <- function(x){  return (qf(1-x, p-1, n-p))  }    tol <- 10^-7  eval\_g\_ineq <- function(x){  return (f - eval\_fa(x)+tol)  }  lb <- 0  ub <- 1  x0 <- 0.05    local\_opts <- list( "algorithm" = "NLOPT\_LD\_MMA", "xtol\_rel" = 1.0e-15 )  opts <- list( "algorithm"= "NLOPT\_GN\_ISRES",  "xtol\_rel"= 1.0e-15,  "maxeval"= 160000,  "local\_opts" = local\_opts,  "print\_level" = 0 )    res <- nloptr ( x0 = x0,  eval\_f = eval\_fa,  lb = lb,  ub = ub,  eval\_g\_ineq = eval\_g\_ineq,  opts = opts  )    return(floor(res$solution/tol)\*tol)  } |

Підрахунок індексу кореляції для пар досліджуваних змінних

Власне, використаємо попередньо наведену функцію для підрахунку індексу кореляції для кожної з пар досліджуваних змінних:

|  |  |
| --- | --- |
| **Пара змінних** | **Код і скріншот результату** |
| duration, popularity | check\_correlation(songs$duration, songs$popularity) |
|  |
| duration, tempo | check\_correlation(songs$duration, songs$tempo) |
|  |
| popularity, tempo | check\_correlation(songs$popularity, songs$tempo) |
|  |

Підрахунок максимального рівня значущості при якому парний статистичний зв'язок не є значимим

Власне, використаємо попередньо наведену функцію для підрахунку максимального рівня значущості при якому парний статистичний зв'язок для кожної з пар досліджуваних змінних не є значимим:

|  |  |  |
| --- | --- | --- |
| **Пара змінних** | **Код і скріншот результату** | **αmax** |
| duration, popularity | max\_alpha(songs$duration, songs$popularity) | 0.532491 |
|  |
| duration, tempo | max\_alpha(songs$duration, songs$tempo) | 0.000104 |
|  |
| popularity, tempo | max\_alpha(songs$popularity, songs$tempo) | 0.000717 |
|  |

Перевірка правильності підрахунку максимального рівня значущості

Перевіримо, для прикладу, значення рівня значущості, отримане для змінних duration та tempo. Перевіримо значимість зв’язку для рівня значущості 0.0001041, та дещо більшого:

|  |
| --- |
| check\_correlation(songs$duration, songs$tempo, 0.0001041) |
|  |
| check\_correlation(songs$duration, songs$tempo, 0.0001042) |
|  |

В першому випалку зв’язок не є істотним, в другому – є. Отже, максимальний рівень значущості при якому парний статистичний зв'язок не є значимим підраховано правильно.

Впорядкована послідовність пар змінних у порядку спадання істотності статистичного зв'язку між ними

Використаємо підрахований індекс кореляції і отримаємо наступну послідовність:

1. duration, tempo (0.0022)
2. popularity, tempo (0.0017)
3. duration, popularity (0.00006)

**Висновки по кореляційному аналізу парних статистичних зв'язків**

На основі підрахованих в цьому розділі характеристиках статистичного

зв'язку можемо зробити наступні висновки:

* Зв’язок між довжиною треку та його популярністю відсутній
* Існує зв’язок між довжиною треку та його темпом
* Існує зв’язок між популярністю треку та його темпом

Аналіз істотності множинних статистичних зв'язків

Функція підрахунку індексу кореляції та перевірки його на істотність

Для зручного аналізу змінних, створимо єдину функцію, яка буде вираховувати індекс кореляції, та перевіряти індекс кореляції на значимість.

Код функції наступний:

|  |
| --- |
| check\_correlation <- function(x, a, b, alpha=0.05){  n <- length(x)  p <- 3  indcor <- summary(lm(x ~ a + b))$r.squared  print(paste("Індекс кореляції:", indcor))    if(indcor == 0){  print("Зв'язку між змінними немає.")  } else if (indcor == 1){  print("Зв'язок між змінними функціональний.")  } else {    # Перевірка гіпотези I=0  f <- (indcor/(1-indcor))\*((n-p)/(p-1))  fa <- qf(1-alpha, p-1, n-p)    if(f < fa){  print("Зв'язок між змінними не є істотним.")  print("Прийнято гіпотезу про незначимість індексу кореляції:")  print(paste(f, "<", fa))  } else {  print("Зв'язок між змінними є істотним.")  print("Гіпотезу про незначимість індексу кореляції не прийнято:")  print(paste(f, ">=", fa))  }    }  return(indcor)  } |

Функція підрахунку максимального рівня значущості при якому відповідний множинний статистичний зв'язок не є значимим

Створимо функцію, що повертатиме максимальний рівень значущості при якому відповідний множинний статистичний зв'язок не є значимим. Скористаємось функцією оптимізації з бібліотеки *nloptr*, яка повертає максимальний рівень значущості, при якому виконується нерівність .

|  |
| --- |
| install.packages("nloptr")  library("nloptr")  max\_alpha <- function(x, a, b){  n <- length(x)  p <- 3  indcor <- summary(lm(x ~ a + b))$r.squared  f <- (indcor/(1-indcor))\*((n-p)/(p-1))    eval\_fa <- function(x){  return (qf(1-x, p-1, n-p))  }    tol <- 10^-7  eval\_g\_ineq <- function(x){  return (f - eval\_fa(x)+tol)  }    lb <- 0  ub <- 1  x0 <- 0.05    local\_opts <- list( "algorithm" = "NLOPT\_LD\_MMA", "xtol\_rel" = 1.0e-15 )  opts <- list( "algorithm"= "NLOPT\_GN\_ISRES",  "xtol\_rel"= 1.0e-15,  "maxeval"= 160000,  "local\_opts" = local\_opts,  "print\_level" = 0 )    res <- nloptr ( x0 = x0,  eval\_f = eval\_fa,  lb = lb,  ub = ub,  eval\_g\_ineq = eval\_g\_ineq,  opts = opts  )    return(floor(res$solution/tol)\*tol)  } |

Підрахунок індексу кореляції для досліджуваних змінних

Власне, використаємо попередньо наведену функцію для підрахунку індексу кореляції для кожної з досліджуваних змінних як для залежної:

|  |  |
| --- | --- |
| **Змінні** | **Код і скріншот результату** |
| duration від (popularity + tempo) | check\_correlation(songs$duration, songs$popularity, songs$tempo) |
|  |
| tempo від (duration + popularity) | check\_correlation(songs$tempo, songs$duration, songs$popularity) |
|  |
| popularity від (tempo + duration) | check\_correlation(songs$popularity, songs$tempo, songs$duration) |
|  |

Підрахунок максимального рівня значущості при якому множинний статистичний зв'язок не є значимим

Власне, використаємо попередньо наведену функцію для підрахунку максимального рівня значущості при якому множинний статистичний зв'язок не є значимим:

|  |  |  |
| --- | --- | --- |
| **Пара змінних** | **Код і скріншот результату** | **αmax** |
| duration від (popularity + tempo) | max\_alpha(songs$duration, songs$popularity, songs$tempo) | 0.000481 |
|  |
| tempo від (duration + popularity) | max\_alpha(songs$tempo, songs$duration, songs$popularity) | 0.000001 |
|  |
| popularity від (tempo + duration) | max\_alpha(songs$popularity, songs$tempo, songs$duration) | 0.002938 |
|  |

Перевірка правильності підрахунку максимального рівня значущості

Перевіримо, для прикладу, значення рівня значущості, отримане для змінної popularity як залежної. Перевіримо значимість зв’язку для рівня значущості 0.002938, та дещо більшого:

|  |
| --- |
| check\_correlation(songs$popularity, songs$tempo, songs$duration, 0.002938) |
|  |
| check\_correlation(songs$popularity, songs$tempo, songs$duration, 0.002939) |
|  |

В першому випалку зв’язок не є істотним, в другому – є. Отже, максимальний рівень значущості при якому множинний статистичний зв'язок не є значимим підраховано правильно.

Впорядкована послідовність пар змінних у порядку спадання істотності множинного статистичного зв'язку їх з множиною усіх інших скалярних змінних

Використаємо підрахований індекс кореляції і отримаємо наступну послідовність:

1. tempo від popularity та duration (0.0039)
2. duration від tempo та popularity (0.0023)
3. popularity від tempo та duration (0.0017)

**Висновки по кореляційному аналізу множинних статистичних зв'язків**

На основі підрахованих в цьому розділі характеристиках статистичного

зв'язку можемо зробити наступні висновки:

* Існує зв’язок між темпом треку та вектором популярності та довжини
* Існує зв’язок між довжиною треку та вектором популярності та темпу
* Існує зв’язок між популярністю треку та вектором темпу та довжини

Підсумок по лабораторній роботі

У лабораторній роботі було проведено кореляційний аналіз трьох змінних.

Проведено аналіз істотності парних статистичних зв'язків для усіх пар змінних.

Проведено аналіз істотності множинних статистичних зв'язків між кожною обраною в якості залежної скалярною змінною та множиною усіх інших скалярних змінних.

За підсумками для змінних popularity, tempo та duration деякі парні зв’язки між ними та всі множинні зв’язки є істотними.

Список використаних джерел

1. Слабоспицький О. С. Основи кореляційного аналізу даних: Навчальний посібник / О. С. Слабоспицький., 2006. – 59 с.
2. <https://www.statology.org/test-for-normality-in-r/>
3. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/sample.html>
4. <https://en.wikipedia.org/wiki/Coefficient_of_determination>
5. <http://www.r-tutor.com/elementary-statistics/simple-linear-regression/coefficient-determination>
6. <https://towardsdatascience.com/how-to-solve-a-constraint-optimization-problem-in-r-fdf5abee197b>