НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

КАФЕДРА ІНФОРМАТИКИ ТА ПРОГРАМНОЇ ІНЖЕНЕРІЇ

**КУРСОВА РОБОТА**

з дисципліни «Аналіз даних в інформаційних системах»

на тему: «Аналіз впливу деяких факторів на рівень щастя населення країн.»

Студентів 2 курсу ІП-01 групи

Спеціальності: 121

«Інженерія програмного забезпечення»

Заранік Богдан Юрійович

Пашковський Євгеній Сергійович

«ПРИЙНЯВ» з оцінкою

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

доц. Ліхоузова Т.А. / доц. Олійник Ю.О.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Підпис                    Дата

Київ - 2022 рік

Національний технічний університет України “КПІ ім. Ігоря Сікорського”

Кафедра інформатики та програмної інженерії

Дисципліна Аналіз даних в інформаційно-управляючих системах

Спеціальність 121 "Інженерія програмного забезпечення"

Курс 2 Група ІП-01 Семестр 4

**ЗАВДАННЯ**

**на курсову роботу студентів**

|  |
| --- |
| Зараніка Богдана Юрійовича та Пашковського Євгенія Сергійовича |

|  |
| --- |
| 1.Тема: Аналіз впливу деяких факторів на рівень щастя населення країн. |
|  |

|  |  |
| --- | --- |
| 2.Строк здачі студентом закінченої роботи | 19.06.2022 |

|  |  |
| --- | --- |
| 3. Вхідні дані до роботи | методичні вказівки до курсової робота, обрані дані з сайту |
| <https://www.kaggle.com/datasets/virajkulkarni952/country-development-indicators> | |
| <https://www.kaggle.com/datasets/mayzannilarthein44/world-happiness-report-2015-to-2022> | |
| <https://www.kaggle.com/datasets/jamesvandenberg/renewable-power-generation?select=Country_Consumption_TWH.csv> | |
| <https://www.kaggle.com/datasets/saleh846/causes-of-deaths-worldwide?select=age-between-5-and-14.csv> | |
|  | |

4.Зміст розрахунково-пояснювальної записки (перелік питань, які підлягають розробці)

|  |
| --- |
| 1.Постановка задачі |
| 2.Аналіз предметної області |
| 3.Розробка сховища даних |
| 4.Інтелектуальний аналіз даних |

5.Перелік графічного матеріалу ( з точним зазначенням обов’язкових креслень )

|  |  |
| --- | --- |
|  | |
|  | |
|  | |
|  | |
| 6.Дата видачі завдання | 16.04.2022 |

**КАЛЕНДАРНИЙ ПЛАН**

|  |  |  |  |
| --- | --- | --- | --- |
| **№п/п** | **Назва етапів курсової роботи** | **Термін виконання етапів роботи** | **Підписи керівника, студента** |
| **1.** | **Отримання теми курсової роботи** | **16.04.2022** |  |
| **2.** | **Визначення зовнішніх джерел даних** | **20.04.2022** |  |
| **3.** | **Пошук та вивчення літератури з питань курсової роботи** | **25.04.2022** |  |
| **4.** | **Розробка моделі сховища даних** | **01.05.2022** |  |
| **5.** | **Розробка ETL процесів** | **15.05.2022** |  |
| **6.** | **Обґрунтування методів інтелектуального аналізу даних** | **20.05.2022** |  |
| **7.** | **Застосування та порівняння ефективності методів інтелектуального аналізу даних** | **25.05.2022** |  |
| **8.** | **Підготовка пояснювальної записки** | **15.06.2022** |  |
| **9.** | **Здача курсової роботи на перевірку** | **19.06.2022** |  |
| **10.** | **Захист курсової роботи** | **21.06.2022** |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Студент |  |  | Заранік Богдан Юрійович |
|  | (підпис) |  | (прізвище, ім’я, по батькові) |

|  |  |  |  |
| --- | --- | --- | --- |
| Студент |  |  | Пашковський Євгеній Сергійович |
|  | (підпис) |  | (прізвище, ім’я, по батькові) |

|  |  |  |  |
| --- | --- | --- | --- |
| Керівник |  |  | доц. Ліхоузова Т.А |
|  | (підпис) |  | (прізвище, ім’я, по батькові) |
| Керівник |  |  | доц. Олійник Ю.О. |
|  | (підпис) |  | (прізвище, ім’я, по батькові) |

"26" червня 2022 р.

## АНОТАЦІЯ

Пояснювальна записка до курсової роботи: 27 сторінок, 26 рисунки, 11 посилань.

Об’єкт дослідження: інтелектуальний аналіз даних.

Предмет дослідження: створення програмного забезпечення, що дозволить аналізувати залежність рівня щастя від деяких параметрів розвитку країн, його прогнозування та тернарна класифікація країн за рівнем розвитку у залежності від вищезгаданих параметрів.

Мета роботи: проектування та реалізація сховища даних, ETL процесів та імплементація програмного забезпечення мовою Python для отримання даних зі сховища та їх подальшого аналізу, прогнозування та класифікації.

Курсова робота включає в себе: опис проектування, створення та заповнення сховища даних згідно з темою завдання за допомогою використання бібліотек мови Python та скриптів SQL, опис створення програмного забезпечення для інтелектуального аналізу даних, їх графічного представлення у вигляді графіків та гістограм та прогнозування за допомогою математичних моделей.

МОДЕЛЬ ПРОГНОЗУВАННЯ, КЛАСИФІКАЦІЯ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ETL ПРОЦЕСИ.

## ЗМІСТ

[АНОТАЦІЯ 3](#_Toc24368)

[ЗМІСТ 4](#_Toc26000)

[ВСТУП 5](#_Toc21063)

[1. ПОСТАНОВКА ЗАДАЧІ 6](#_Toc395)

[2. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ 7](#_Toc29883)

[3. РОЗРОБКА СХОВИЩА ДАНИХ 9](#_Toc13477)

[3.1. Розробка ETL процесів 9](#_Toc4903)

[3.1.1. ETL для датасету World happiness report 2015-2022 9](#_Toc7379)

[3.1.2. ETL для датасету Deaths Reasons. 10](#_Toc20585)

[3.1.3. ETL для Country Consumption 11](#_Toc17228)

[3.1.4. Скрипт mainETL.py 12](#_Toc869)

[3.2. Створення сховища даних 13](#_Toc29635)

[3.3. Завантаження даних до сховища 15](#_Toc24537)

[4. ВІДБІР ТА АНАЛІЗ ДАНИХ 17](#_Toc13937)

[4.1. Візуалізація даних 17](#_Toc10955)

[4.2. Створення представлень бази даних 19](#_Toc19276)

[4.3. Підготовка даних до аналізу 20](#_Toc32591)

[5. ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ 24](#_Toc14369)

[5.1. Обґрунтування алгоритмів для побудови регресійної моделі 24](#_Toc3782)

[5.2. Побудова і тренування моделі 25](#_Toc18549)

[5.3. Валідація моделі 26](#_Toc24511)

[5.4. Візуалізація результатів регресійного аналізу 29](#_Toc5451)

[ПЕРЕЛІК ПОСИЛАНЬ 32](#_Toc26906)

[ДОДАТОК А ТЕКСТИ ПРОГРАМНОГО КОДУ 33](#_Toc7238)

## 

## ВСТУП

Питання рівня щастя населення має першочергове значення у сучасному світі, що розвивається з величезною швидкістю. Від нього залежить рівень життя населення, темпи економічного розвитку країни, привабливість країни для спеціалістів сучасних професій із-за кордону, інвестиційна привабливість та інші.

Цей параметр є достатньо відносним, проте навіть його можна оцінити за певною шкалою. На рівень щастя населення впливають багато чинників, перш за все - економічні та соціальні.

Міжнародний індекс щастя (англ. *Happy Planet Index*) являє собою індекс, що відображає добробут людей та стан навколишнього середовища в різних країнах світу. Головне завдання індексу відобразити «реальний» добробут націй. Для порівняння рівня життя в різних країнах використовується значення ВВП на душу населення або ІЛР, але ці індекси не завжди можуть відобразити реальний стан речей. Зокрема порівняння значення ВВП на душу населення вважається недоречним, оскільки кінцева мета більшості людей не бути багатими, а бути щасливими та здоровими.

У нашій роботі ми задалися питанням, які саме чинники впливають на рівень щастя людей суттєво, а якими можна знехтувати.

Також, як відомо, країни поділяють на “розвинені”, “ті, що розвиваються” та “слабо розвинені”. За даними вибірок ми також маємо на меті класифікувати країни за цими трьома типами, оскільки дуже вірогідно, що параметри, що сильно впливають на рівень щастя населення країн, також будуть суттєво вливати на те, до якого класу розвиненості належить та чи інша країна. Отже, цю гіпотезу нам і потрібно перевірити шляхом створення математичної моделі класифікації даних.

## ПОСТАНОВКА ЗАДАЧІ

Мета нашого дослідження - розробка ПО, що дозволяє виокремити із переліку параметрів, що впливають на рівень щастя населення, ті, що впливають суттєво, та розробити математичну модель прогнозування рівня щастя населення за даними конкретними параметрами. Також метою нашого дослідження є розробка моделі класифікації країн за рівнем розвитку. Дані математичні моделі можуть суттєво допомогти економістам у перевірці своїх припущень щодо встановленого рівня щастя населення для певної країни та віднесення її до певного класу за рівнем розвиненості.

Результатом роботи алгоритму передбачення рівня щастя населення має бути певне дійсне число, що диференціює країну серед інших за вищезгаданим параметром.

Результатом роботи алгоритму класифікації країни за рівнем розвитку має бути один із трьох класів: “розвинена”, “та, що розвивається” та “слабо розвинена” - передбачуваний рівень розвитку даної країни.

## АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

На нашу думку, впливати на рівень щастя населення та рівень розвитку країни можуть такі чинники:

* ВВП на душу населення
* Рівень свободи
* Рівень довіри населення владі
* Щедрість населення
* Розповсюдженість наркотичних засобів
* Захворюваність на на деякі види захворювань
* Соціальна підтримка зі сторони держави
* Споживання електроенергії на душу населення
* Очікувана середня тривалість життя
* Площа країни проживання
* Загальна смертність на 10000 населення

Даний список параметрів, що впливають на рівень щастя населення, є неповним, оскільки на нього впливає безліч чинників, але досліджувати шукану залежніть ми будемо саме за ним, адже на нашу думку у списку присутня більшість параметрів, що так чи інакше складають оцінку рівня щастя.

У програмній системі буде реалізовано наступну функціональність, що включає в себе:

* створення ETL процесів для завантаження даних;
* створення датасету зі сховища у вигляді csv-файлу;
* графічне відображення отриманих результатів та їх аналіз.
* інтелектуальний аналіз даних;
* використання регресійних моделей прогнозування;
* використання математичних моделей класифікації даних(навчанні із вчителем);

Нами було знайдено 4 датасети.

1. World happiness report 2015-2022.

Описує велику кількість показників, що впливають на рівень щастя населення різних країн по роках 2015-2022 та сам рівень щастя.

1. Country Electricity Consumption

Описує параметр споживання електроенергії певної країни певного року .

1. Deaths Reasons

Описує причини смертності населення певних країн по роках.

1. Population By Country

Містить демографічний аналіз певних країн. Параметри такі як густина населення, кількість населення та інші.

## РОЗРОБКА СХОВИЩА ДАНИХ

### Розробка ETL процесів

Після завантаження датасетів їх потрібно ретельно підготувати до завантаження до сховища даних для подальшого зручного інтелектуального аналізу. Для цього нами було створено низку скриптів на мові Python, що перетворюють “сирі” дані у єдиний зручний формат для обробки.

На першому лістингу зображено підключення необхідних бібліотек для перетворення даних та встановлення деяких конфігурації консольного виводу.

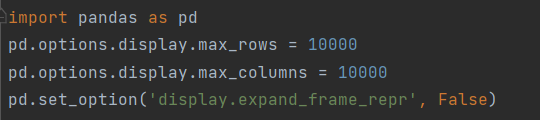


Рисунок 3.1 - Завантаження та імпорт бібліотек

Далі було створено функцію, яка приводить колонку рядкового типу до типу float, при цьому перетворюючи розділювач для дійсних чисел з коми на крапку.

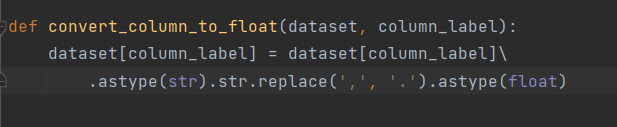


Рисунок 3.2 - Функція-конвертер

#### ETL для датасету World happiness report 2015-2022

Далі потрібно перетворити деякі рядки датасету за допомогою вищезгаданої функції, попередньо завантаживши потрібний датасет до оперативної пам’яті комп’ютера. Також варто помітити, що деякі країни мають зірочку у кінці назви, тому дану неточність потрібно виправити, що і було зроблено.

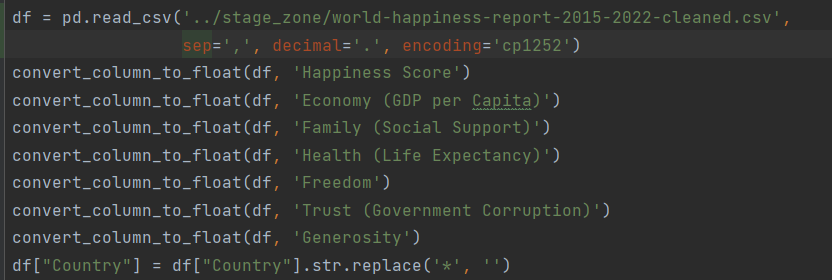


Рисунок 3.3 - Перетворення датасету World happiness report 2015-2022

І насамкінець видалимо непотрібну колонку та збережемо виправлений датасет.

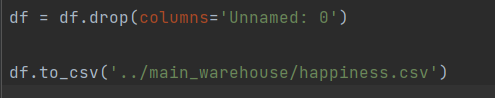


Рисунок 3.4 - Збереження датасету World happiness report 2015-2022

#### ETL для датасету Deaths Reasons.

Датасет Deaths Reasons складений із декількох частин, які потрібно об’єднати. Для цього візьмемо датасети для вікових категорій 15-49, 50-69 та 70+ років. Згрупуємо їх за складеним ключем [Country, Year] та просумуємо задля отримання датасету для вікової категорії 15+ років.

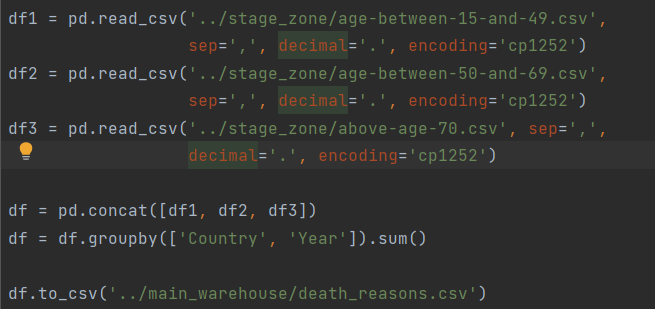


Рисунок 3.5 - Перетворення Deaths Reasons

#### ETL для Country Consumption

Оскільки у даному датасеті колонками є рік та країни(тобто “таблиця широкого формату”), то варто для більш зручного аналізу даних перетворити її у довгий формат за допомогою ф-ції melt. Вона залишить колонку “Year”, а всі колонки із назвами країн потраплять у нову колонку із назвою “Country” як значення. Значення, що були на перехресті конкретного року та у колонці певної країни, потраплять у нову колонку із назвою “Consumption”. Наприклад, наглядно зміна колонок була такою: були колонки Year, China, England…, USA…; стали колонки Year, Country, Consumption.

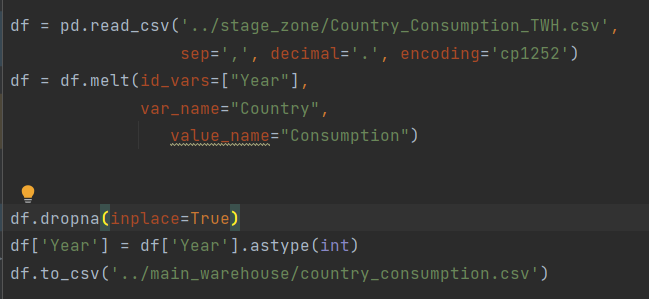


Рисунок 3.5 - Перетворення Country Consumption

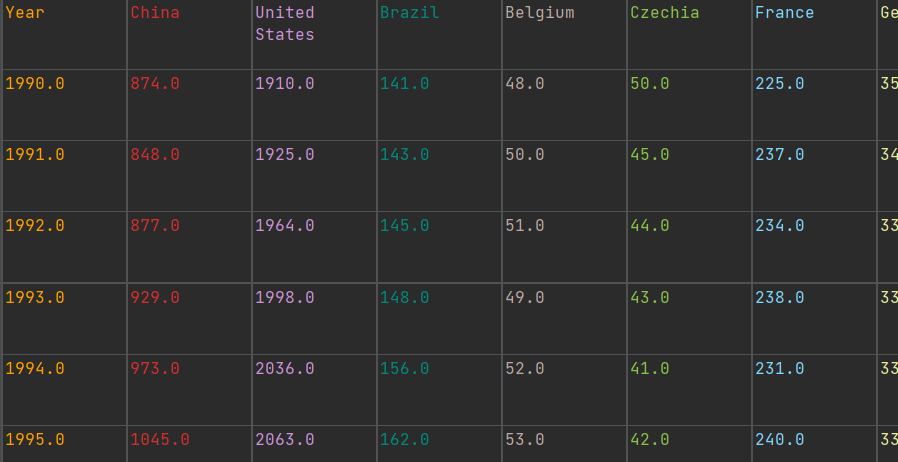


Рисунок 3.6 - До перетворення

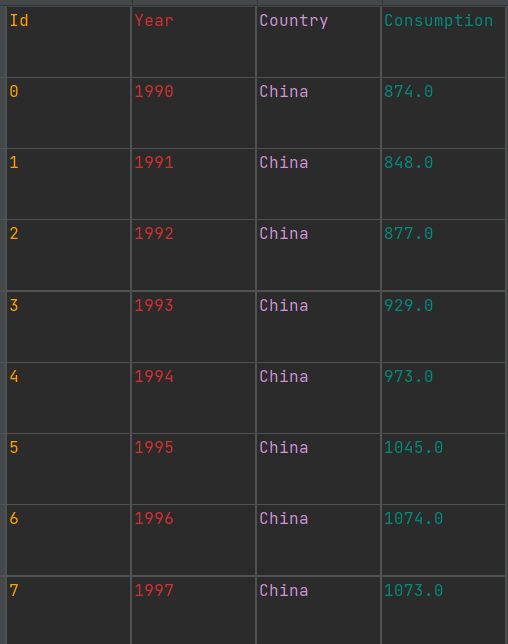


Рисунок 3.7 - Після перетворення

#### Скрипт mainETL.py

Для зручності запуску ETL процесів було створено скрипт, що автоматично запускає усі 3 ETL скрипта послідовно.

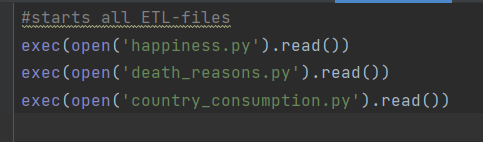


Рисунок 3.8 - Головний ETL скрипт

### Створення сховища даних

Сховище даних складається з двох фактових таблиць та трьох вимірів. Опис таблиць бази даних наведені у наступній Таблиці 1.

|  |  |
| --- | --- |
| Назва таблиці | Семантичне значення таблиці |
| happiness\_report | фактова таблиця, звіт про фактори, що потенційно можуть впливати на рівень щастя та колонка із рівнем щастя |
| death\_report | фактова таблиця, звіт про смертність населення та її причини |
| date | часовий вимір, представлення року |
| country | вимір, представлення країни |
| death\_reason | вимір, представлення причин смертності |

Таблиця 1 - Опис таблиць бази даних

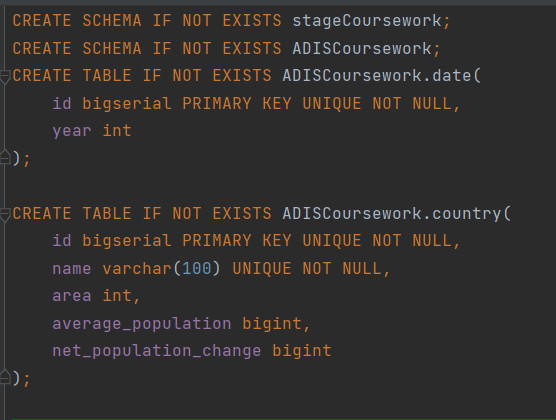


Рисунок 3.9 - Створення таблиць date і country

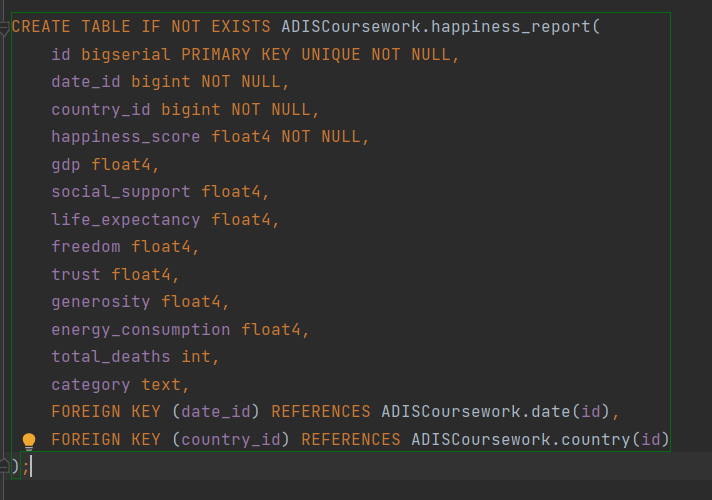


Рисунок 3.9 - Створення таблиці happiness\_report

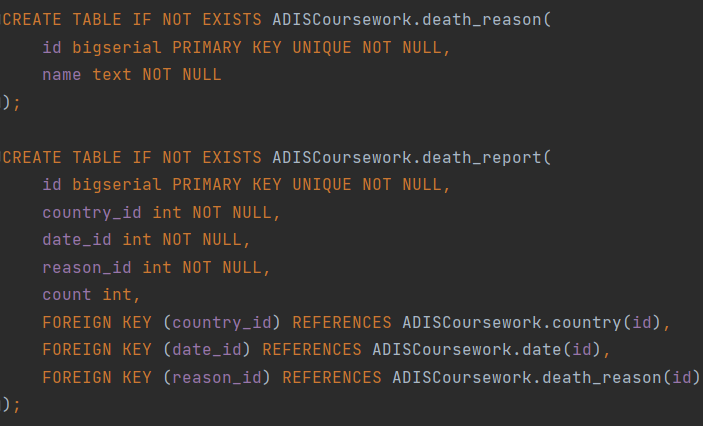


Рисунок 3.10 - Створення таблиць death\_report та death\_reason

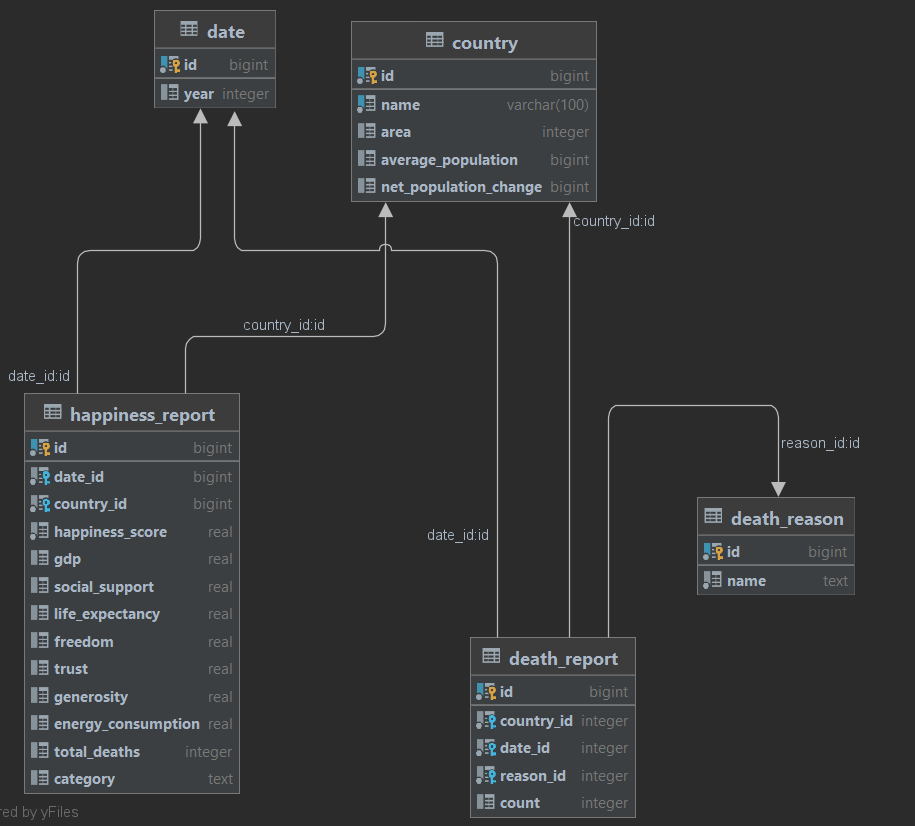


Рисунок 3.10 - Схема сховища даних

### Завантаження даних до сховища

Наступним кроком завантажимо дані зі stage-зони до сховища даних, що було створено у попередньому пункті. Для цього використаємо наступні SQL-скрипти із вкладеними запитами.

Завантажимо дані до таблиці countries.

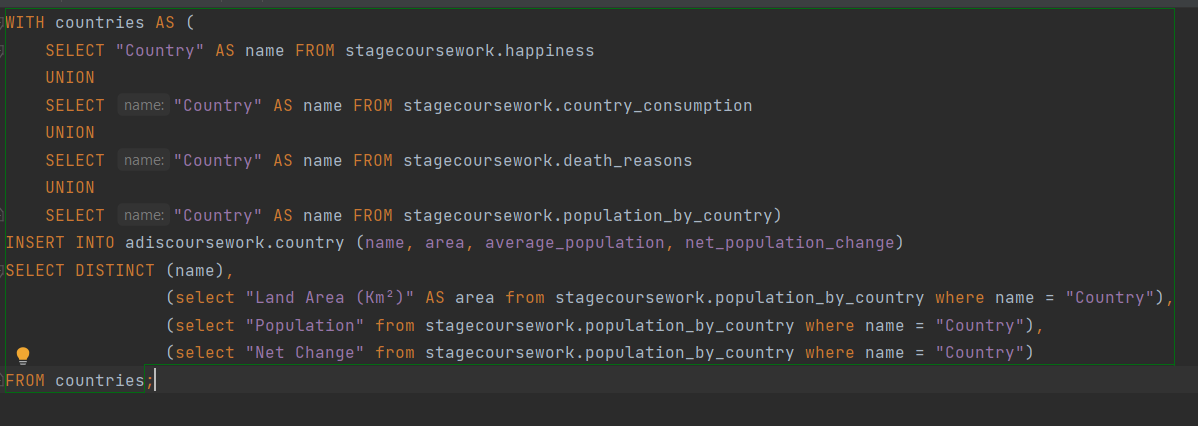


Рисунок 3.11 - Завантаження даних до таблиці countries

Завантажимо дані до таблиці years.

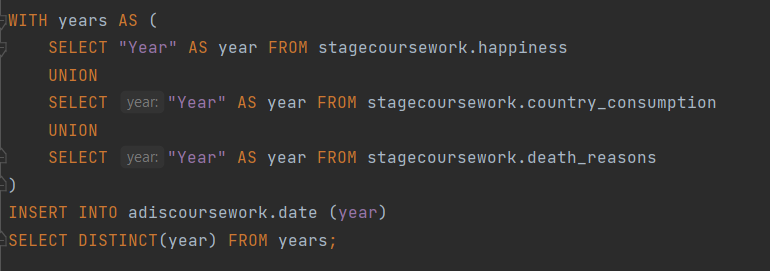


Рисунок 3.11 - Завантаження даних до таблиці years

Завантажимо дані до таблиці happiness\_report, використовуючи вкладені запити.

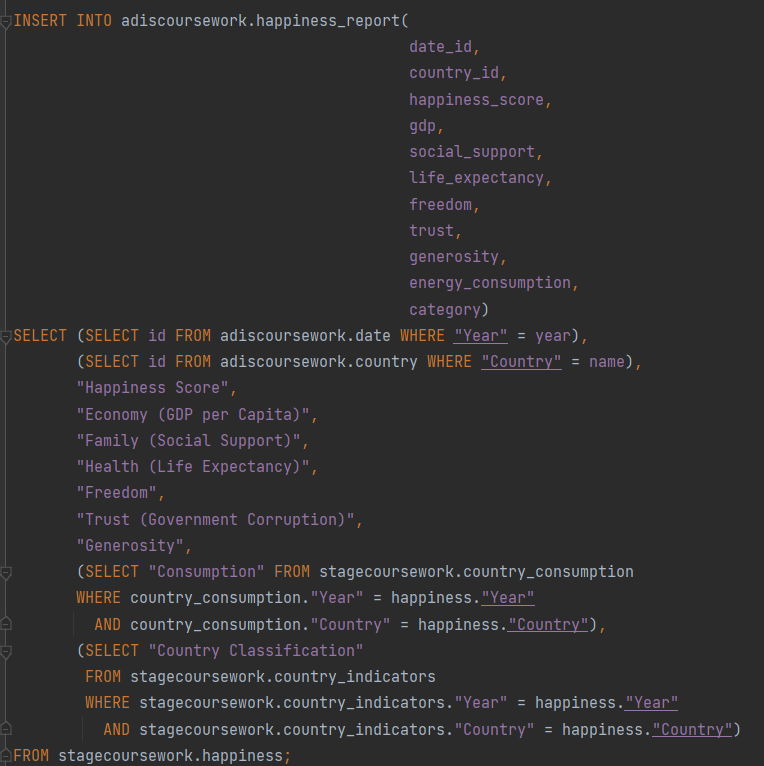


Рисунок 3.11 - Завантаження даних до таблиці happiness\_report

Завантажимо дані до таблиць death\_reason і death\_report, використовуючи вкладені запити.

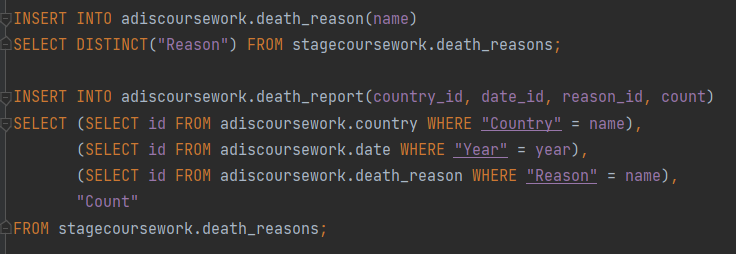


Рисунок 3.11 - Завантаження даних до таблиці death\_reason і death\_report

Отже, результатом роботи над цим розділом стало створення сховища даних і також його заповнення даними з датасетів.

## ВІДБІР ТА АНАЛІЗ ДАНИХ

### Візуалізація даних

Для кращого розуміння, давайте зробимо деяку візуалізацію нашого датасету. Ми знаємо, що наші дані містить колонки, наприклад, gdp, freedom, trust тож побудуємо гістограму, яка показує кількісний розподіл gdp, freedom, trust.

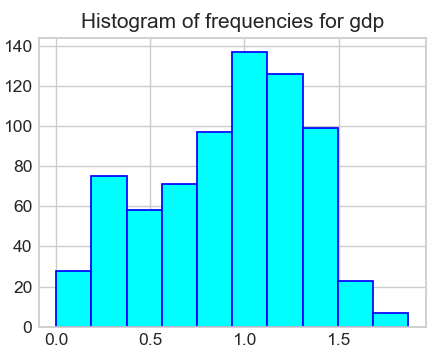


Рисунок 3.13 - Побудова гістограми для gdp

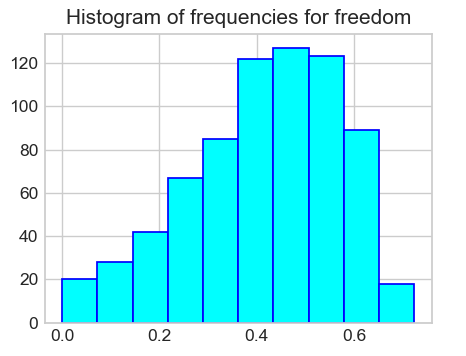


Рисунок 3.14 - Побудова гістограми для freedom

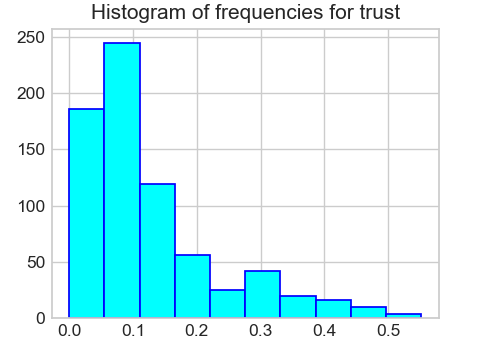


Рисунок 3.14 - Побудова гістограми для trust

Для більшої наглядності продемонструємо також діаграми розмаху деяких даних по колонках:

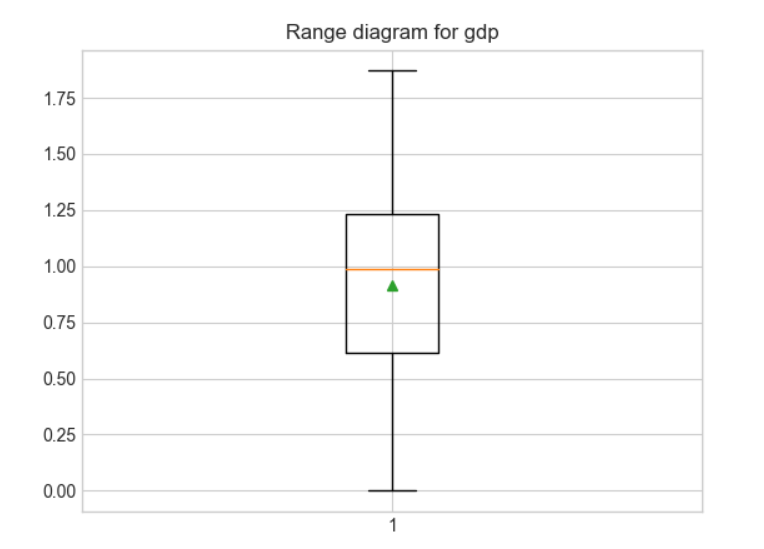


Рисунок 3.14 - Побудова діаграми розмаху для gdp

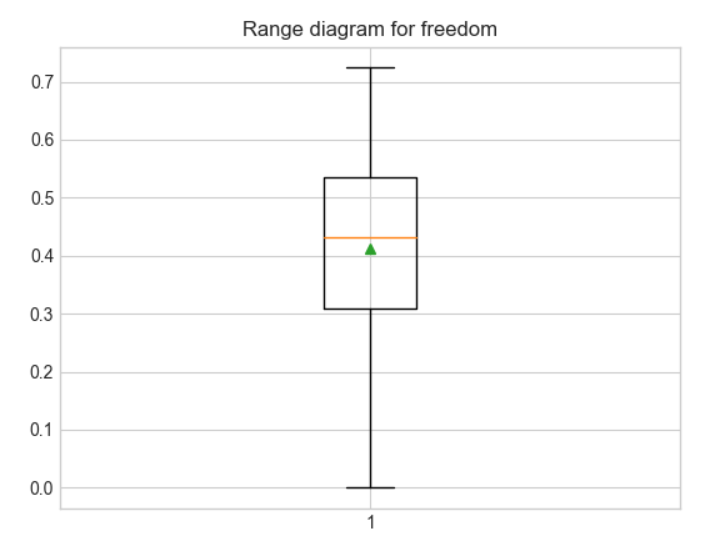


Рисунок 3.14 - Побудова діаграми розмаху для freedom

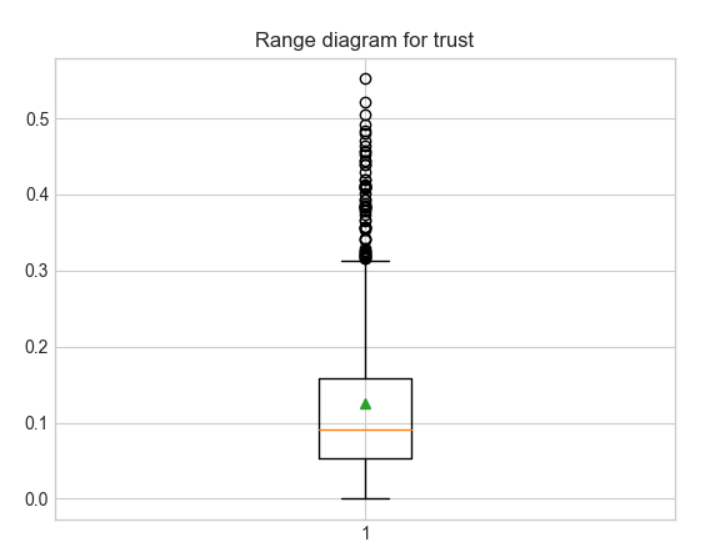


Рисунок 3.14 - Побудова діаграми розмаху для trust

### Створення представлень бази даних

Для отримання зручних для обробки даних із новоствореного сховища даних створимо кілька представлень засобами СУБД мовою SQL.

Для початку створимо представлення, що описує звіт про рівень щастя населення країн світу протягом 2015-2019 років. У представленні присутня також нормалізація деяких полів. Наприклад, замість номінального показника споживання електроенергії, будемо використовувати нормалізоване значення споживання електроенергії на душу населення.

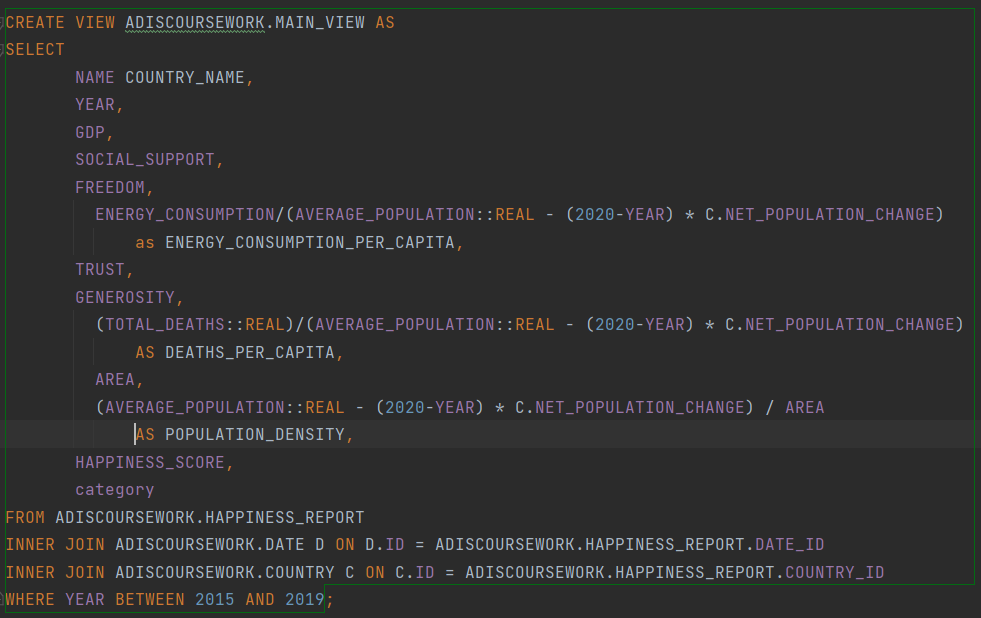


Рисунок 3.14 - Скрипт для створення представлення, що описує звіт про рівень щастя

Далі створимо представлення, що описує вплив різних причин смертності на щастя протягом 2015-2019 років. У представленні присутня також нормалізація деяких полів. Наприклад, замість номінального показника споживання кількості смертей з тієї чи іншої причини, будемо використовувати нормалізоване значення кількості смертей з тієї чи іншої причини на душу населення.

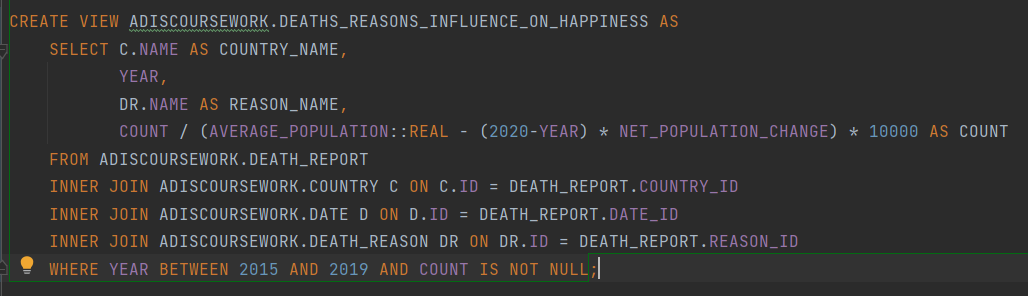


Рисунок 3.14 - Скрипт для створення представлення, що описує вплив різних причин смертності на щастя

### Підготовка даних до аналізу

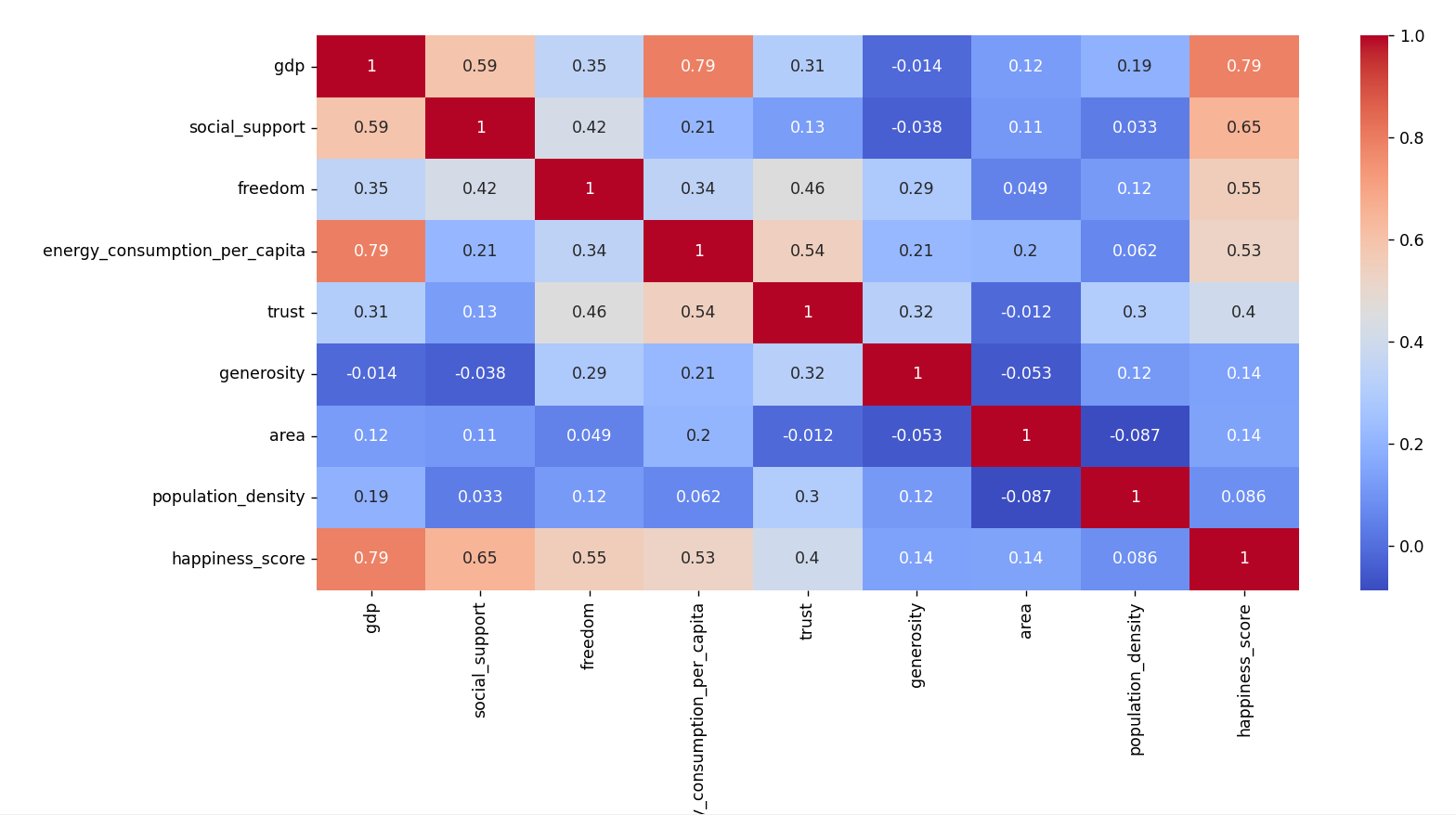


Рисунок 3.14 - Кореляційна матриця до добору параметрів

Оскільки параметри area (кореляція із happiness\_score 0.062), generosity (кореляція із happiness\_score 0.14) і population\_density (кореляція із happiness\_score -0.045) мають низьку значимість, то їх було видалено із аналізу.

Також у зв’язку з мультиколінеарністю було видалено параметр energy\_consumption\_per\_capita (кореляція із gdp 0.79), social\_support(кореляція із gdp 0.59).

Отже, у кінці кінців з представлення main\_view залишись такі параметри(представлена матриця кореляції).

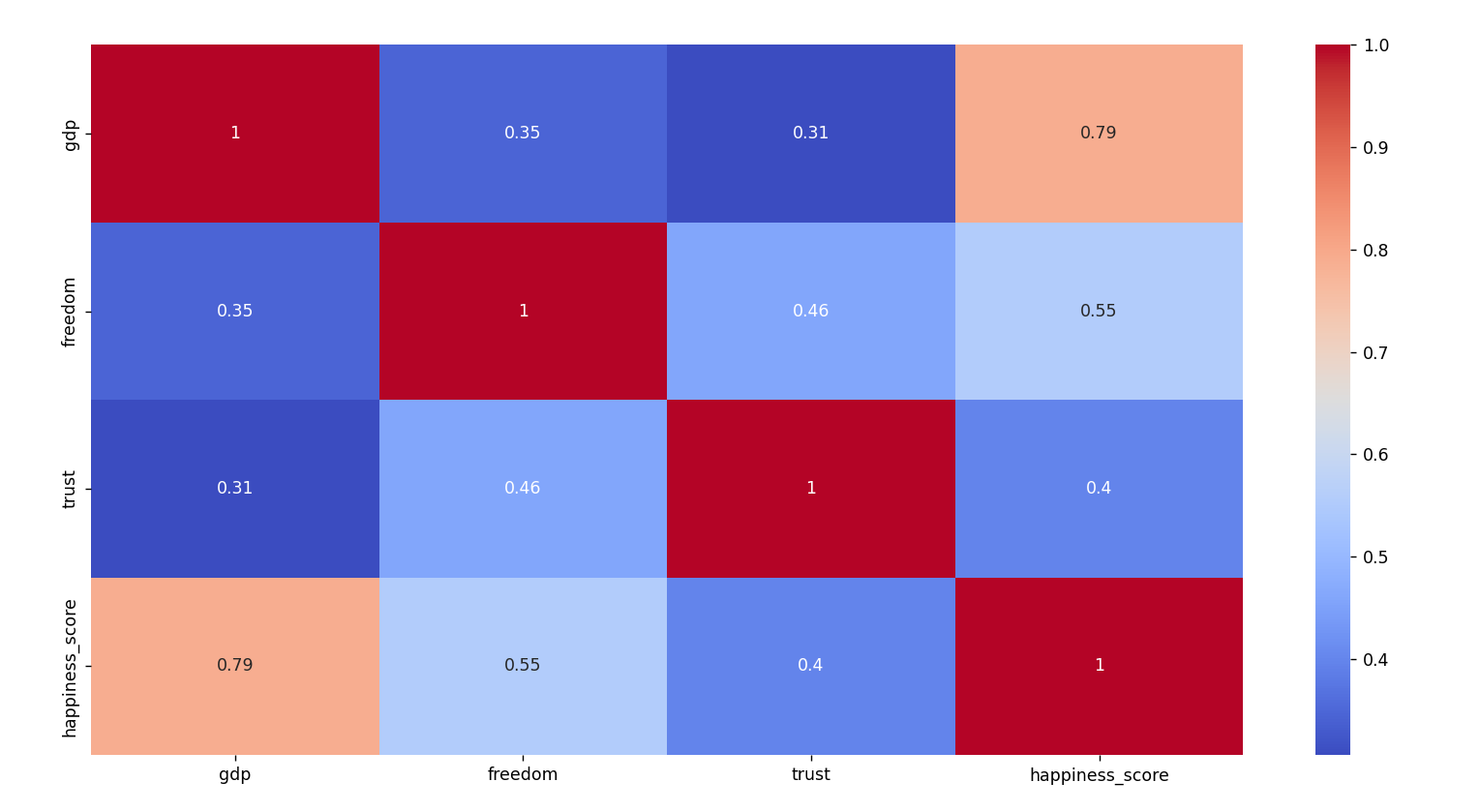


Рисунок 3.14 - Кореляційна матриця після добору параметрів

\* Зауваження: happiness\_score є не предиктором, а результуючим значенням, тому значення значення, наприклад, на перетині happiness\_score та gdp, що рівне 0.79, є не великим значенням кореляції, яке потрібно прибрати, а значення залежності happiness\_score від предиктора gdp.

Розглянемо інше представлення - представлення із причинами смертності. Проаналізувавши таблицю кореляції більш ніж 20 різних причин смертності і їх впливу на рівень щастя населення країн, прийшли до такого висновку, що варто вилучити з розгляду всі параметри, що не задовольняють двома умовам:

* значимість цих параметрів (кореляція з happiness\_score) за модулем не менша за 0.3;
* кореляція між цими параметрами за модулем не більша 0.6;

Отримані параметри “примішаємо” до першого представлення і отримаємо кінцевий результат даних, підготовлених до інтелектуального аналізу.

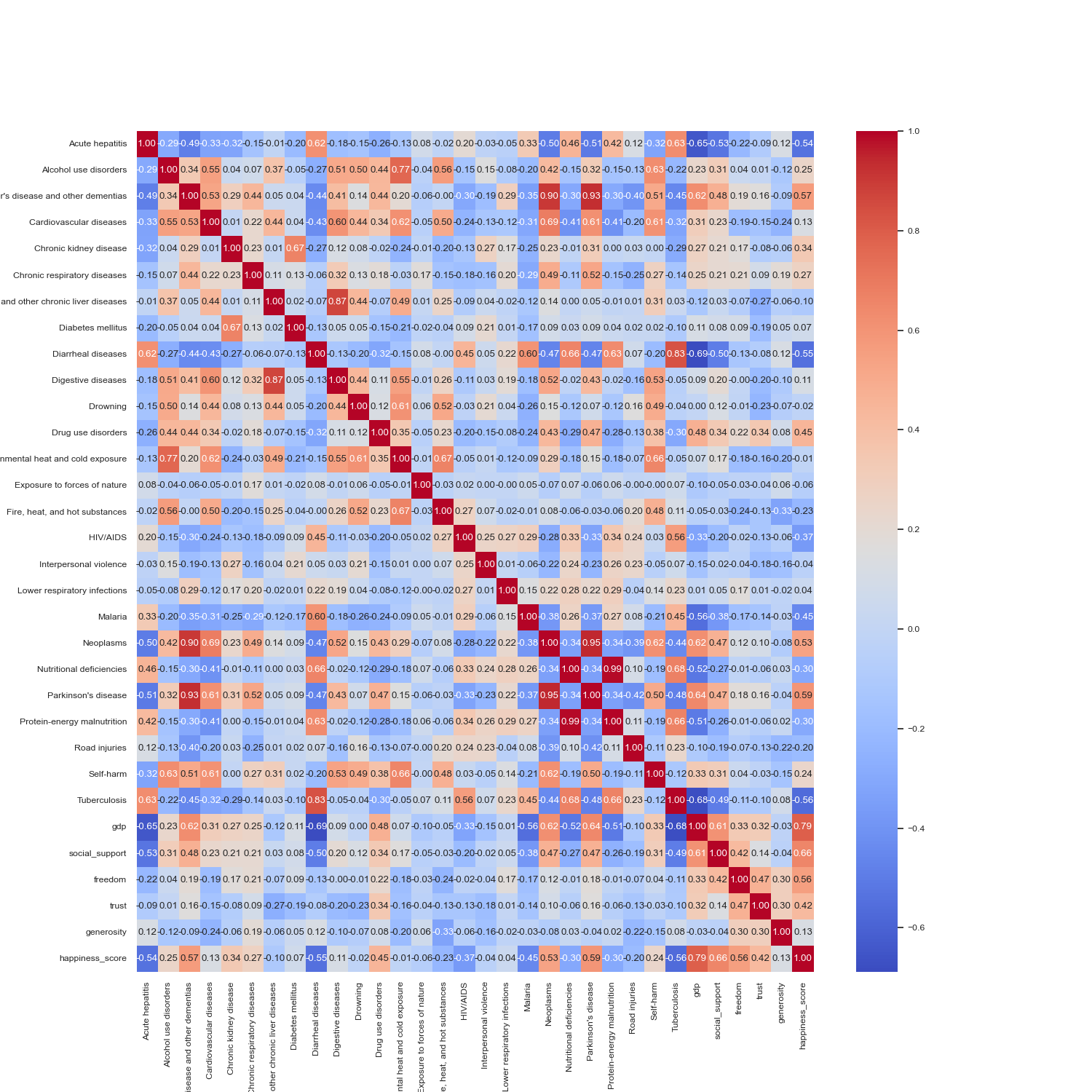


Рисунок 3.14 - Загальна кореляційна матриця до добору

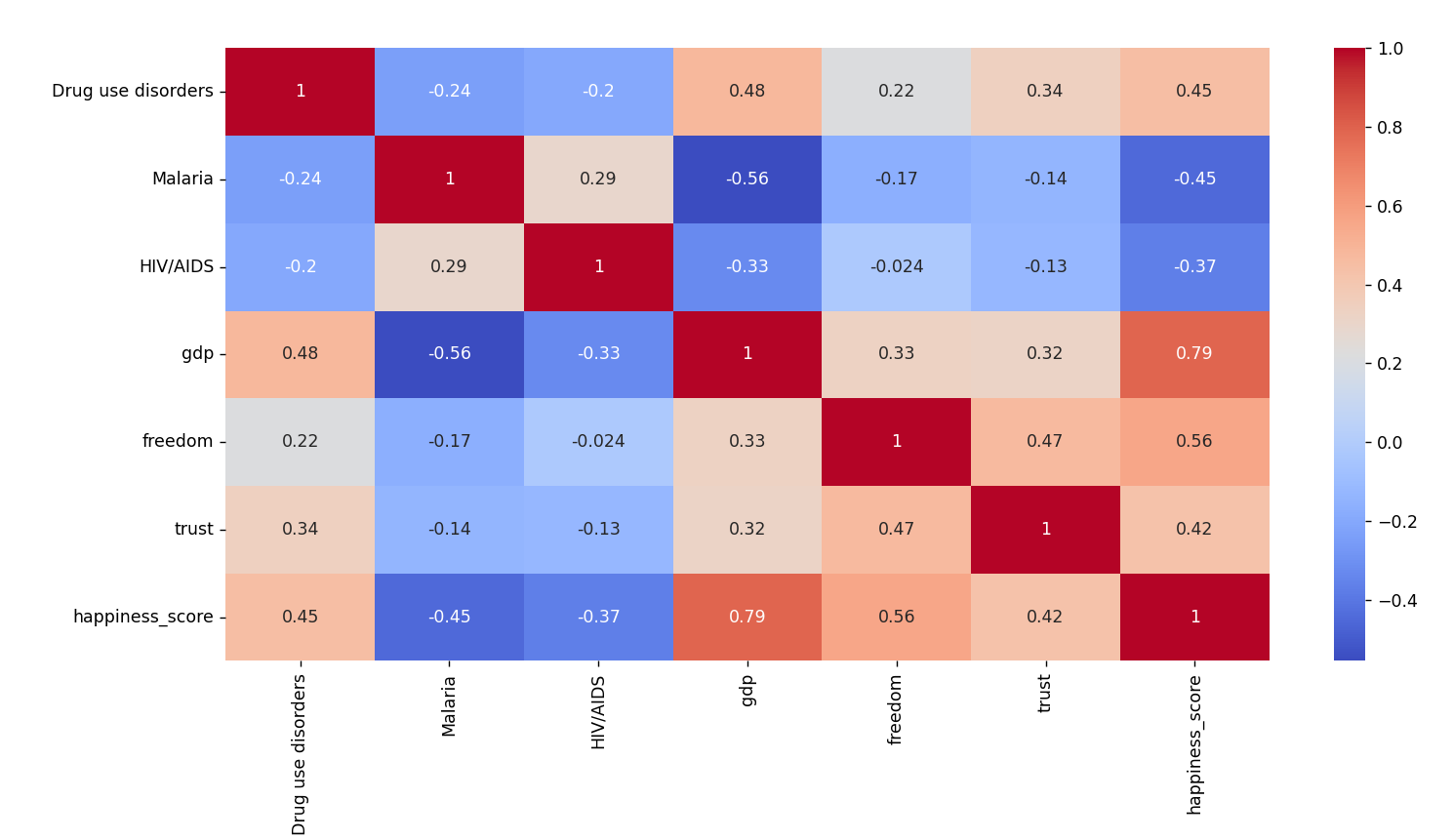


Рисунок 3.14 - Загальна кореляційна матриця після добору

Отже, результатом роботи над цим розділом стали дані, повністю підготовлені до інтелектуального аналізу.

## ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

### Обґрунтування алгоритмів для побудови регресійної моделі

Для побудови регресійної моделі ми використовуємо лінійну, поліноміальну регресії.

Регресійний аналіз – це метод моделювання даних, які вимірюються, та дослідження їх властивостей. Регресійна модель – це функція незалежної величини та коефіцієнтів з включеними випадковими змінними.

Вважають, що залежна змінна описується сумою значень деякої моделі та незалежними змінними. Відповідно до характеру розподілу залежної змінної роблять припущення, які називаються гіпотезою породження даних. Для підтвердження або спростування цієї гіпотези проводяться статистичні тести (аналіз залишків – різниця між значеннями, які спостерігаються, та значеннями, які передбаченні побудованою регресійною моделлю). При цьому вважають, що залежна змінна не містить помилок.

Нами було обрано лінійну та поліноміальну регресії.

***Лінійна регресійна*** модель має наступний вигляд:

|  |  |
| --- | --- |
|  | (4.1) |

де *у* – залежна змінна;

– незалежні змінні;

*u* – випадкова похибка, розподіл якої в загальному випадку залежить від незалежних змінних, але математичне очікування якої рівне нулю.

***Нелінійна регресія*** – окремий випадок регресійного аналізу, в якому розглянутою регресійною моделлю є нелінійна функція, що залежить від параметрів і від однієї або декількох вільних змінних. Відмінність від лінійної регресії полягає тільки в формі зв’язку та методах оцінки параметрів (формула самої регресійної функції. Що призвана оцінювати дані).

Оскільки точність методу лінійної та поліноміальної регресій є достатньою для більшості задач, що виникають у житті, а реалізація відносно простою, то наш вибір є небезпідставним.

Модель завжди є спрощенням реальності, тому вона повинна бути досить проста. З двох моделей, що приблизно однаково відповідають даним, перевагу варто віддати більш простій моделі, що містить, наприклад, менше число пояснюючих змінних.

Оцінити точність регресійної моделі можна за критеріями (коефіцієнт детермінації) і RSE (стандартна похибка залишків), щодасть нам змогу зрозуміти, чи достатньо математична модель описує життєву.

Формула для розрахунку стандартної похибки залишків:

|  |  |
| --- | --- |
|  | (4.2) |

Формула для розрахунку значення коефіцієнту детермінації:

|  |  |
| --- | --- |
| = 1 - | (4.3) |

- де RSS - сума квадратів залишків, TSS - загальна сума квадратів відхилень.

### Побудова і тренування моделі

Для побудови моделі нами було розроблено функцію regression, що одразу генерує, навчає, тестує та оцінює якість регресійної моделі.

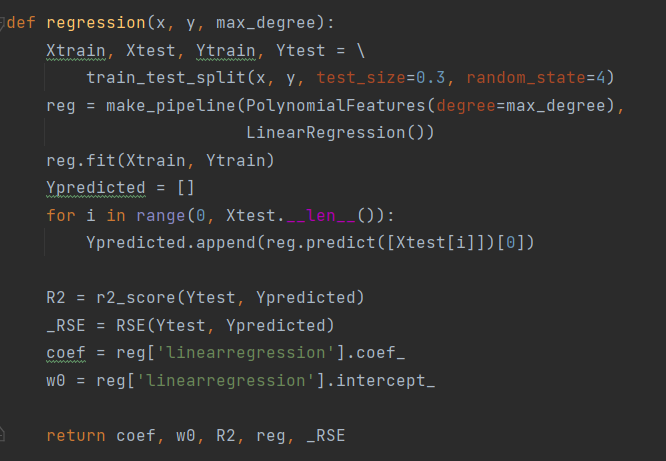


Рисунок 3.14 - Вигляд функції regression

Для побудови моделі спочатку розіб’ємо задану вибірку на навчальну та тестову частини у відношенні 70% до 30% відповідно. Надалі навчальну вибірку використовуватимемо для навчання моделі, а тестову - для перевірки якості моделі на нових для неї даних задля чистоти експерименту. У результаті роботи функції отримали такі змінні:

* Xtrain - предиктори навчальної вибірки
* Ytrain - результати навчальної вибірки
* Xtest - предиктори тестової вибірки
* Ytest - результати тестової вибірки



Рисунок 3.14 - Розбиття вибірки на навчальну та тестову

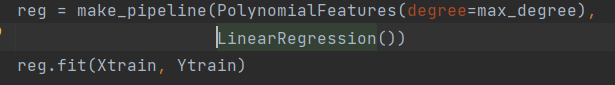


Рисунок 3.14 - Генерація поліному степеня max\_degree та навчання моделі на даних навчальної вибірки

Для кожного X із тестової вибірки Xtest розрахуємо та збережемо передбачене значення за допомогою нашої функції регресії.

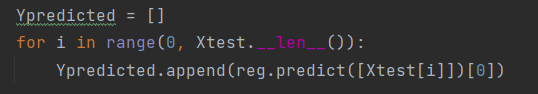


Рисунок 3.14 - Розрахунок та збереження передбачених значень

### Валідація моделі

Під валідацією моделі розуміється оцінка точності прогнозування за вищезгаданими критеріями та RSE.

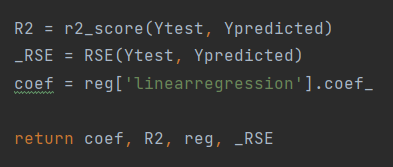


Рисунок 3.14 - Оцінка моделі

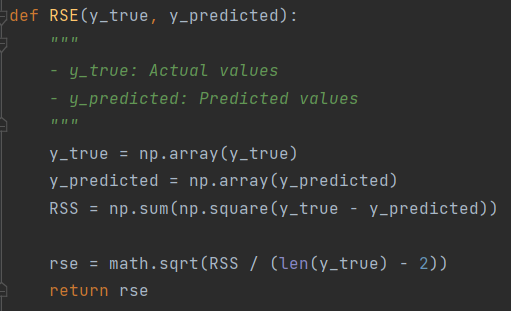


Рисунок 3.14 - Функція для оцінки моделі RSE

Далі нами було розроблено функцію, що розбиває датасет на предиктори і значення та запускає процес побудови моделей для різних степенів для подальшого порівняння між собой та вибору тої, що підходить найбільше.

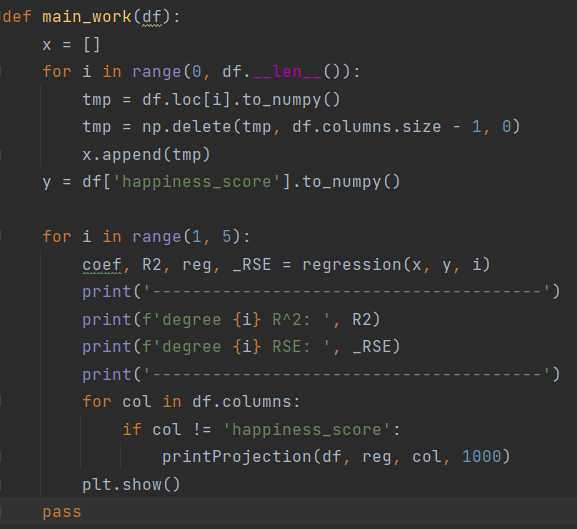
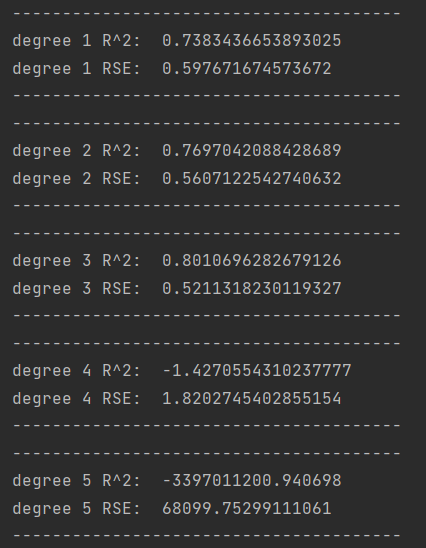


Рисунок 3.14 - Функція main\_work

Рисунок 3.14 - Результат роботи функції main\_work

Із результатів роботи регресії різних степенів очевидно, що поліноміальна регресія степеня 3 має найбільшу оцінку та найменшу оцінку RSE, що вказує на те, що саме ця модель є найбільш точною серед перелічених. Отже, виберемо її.



### Візуалізація результатів регресійного аналізу

Для зручності оцінки адекватності моделі нами біло прийнято рішення створення функції, що зображає двовимірні проекції залежності рівня щастя від предикторів. На графіку буде зображено як емпіричні дані, так і проекцію лінії регресії. Хочеться зауважити, що зображено буде саме проекцію лінії регресії на площину, а не саму лінію, оскільки через велику кількість параметрів візуально зобразити сам графік регресії неможливо.



Рисунок 3.14 - Функція для зображення проекції

Замість усіх предикторів, крім того, що бере участь у проекції, вставляємо нулі та викликаємо функцію передбачення для регресії. Враховуючи те, що усі параметри, що є предикторами, за умовою набувають лише додатніх значень, то такий трюк є цілком припустимим.

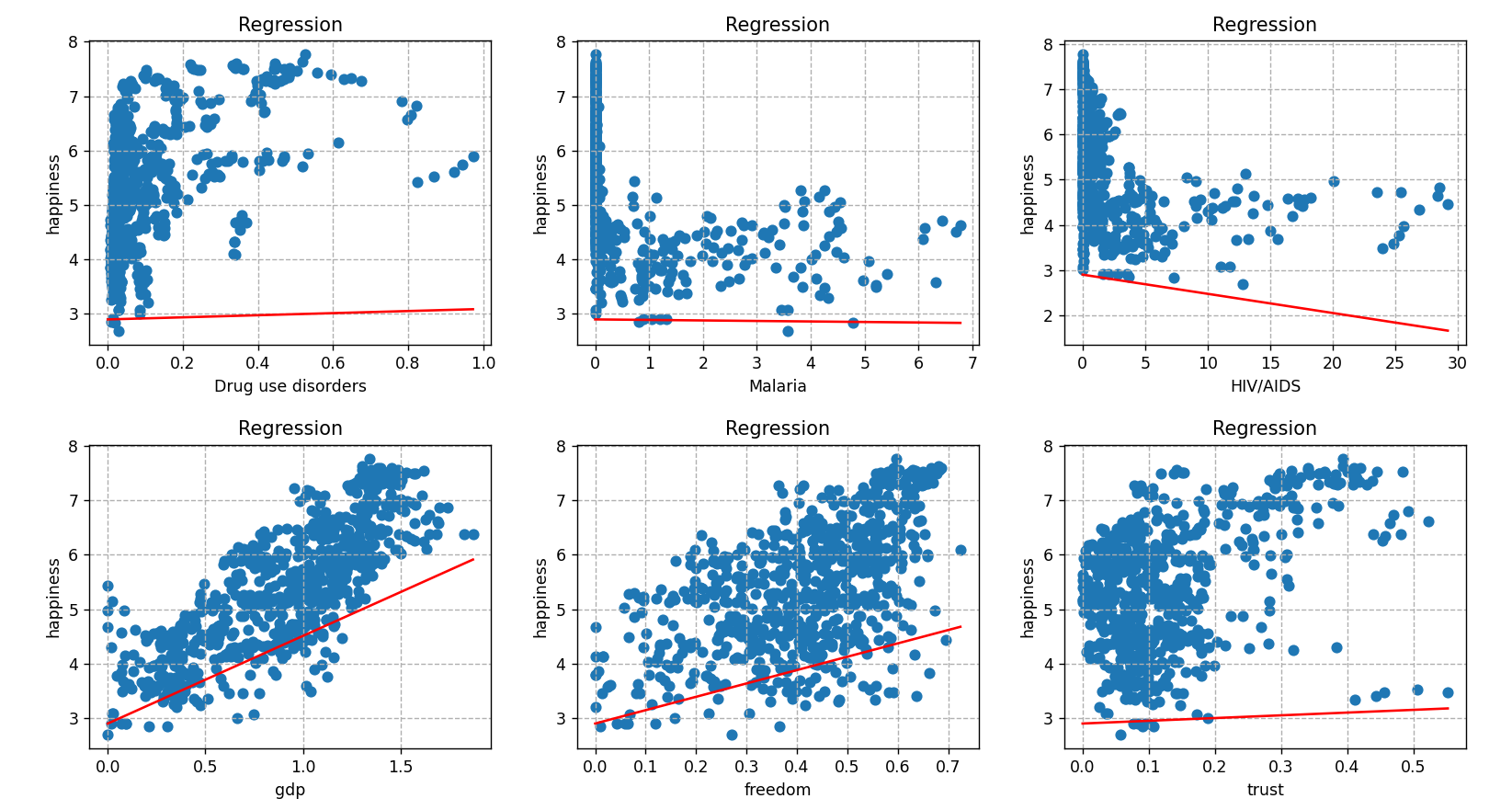


Рисунок 3.14 - Графічне відображення проекції лінійної регресії

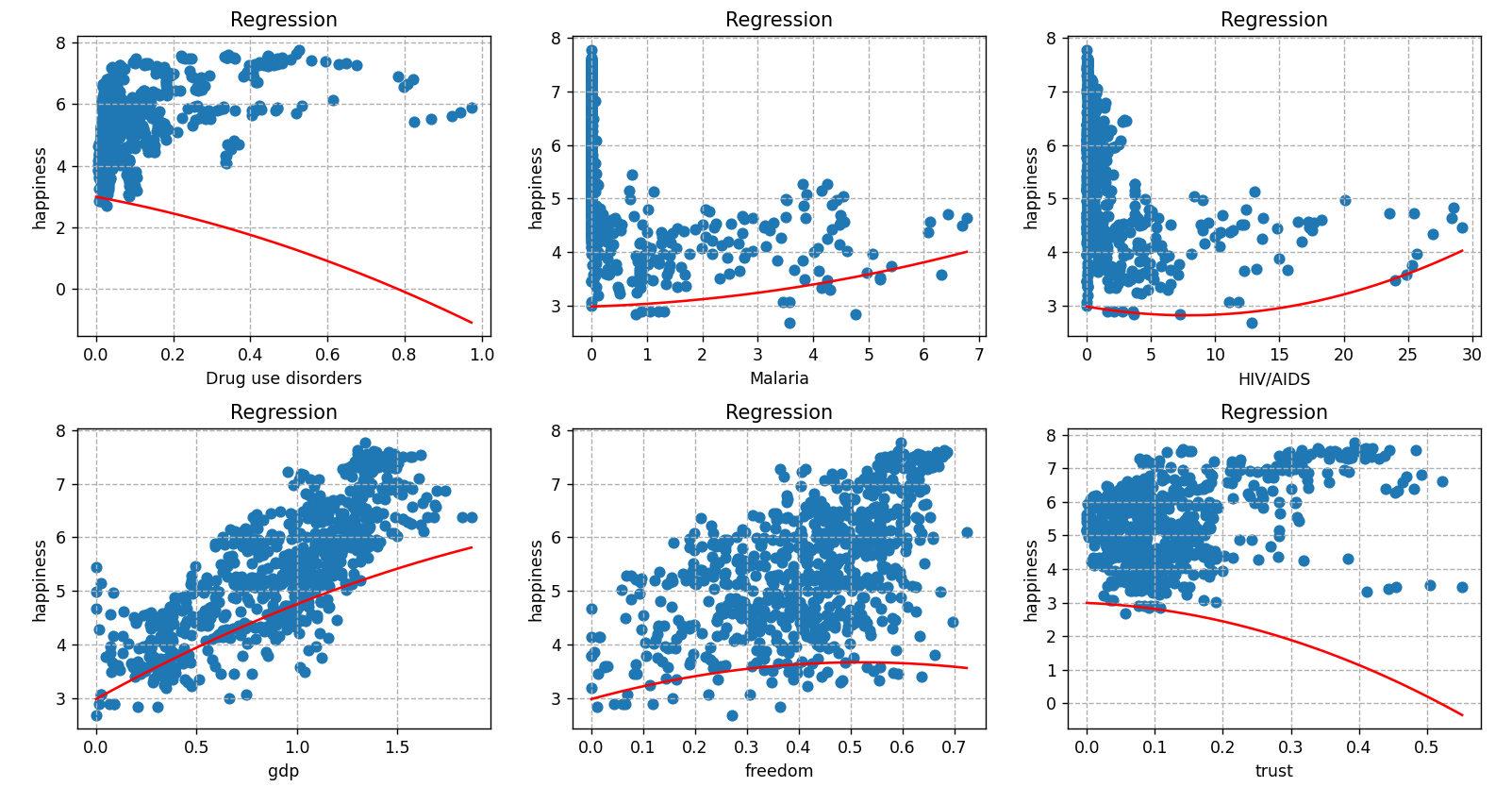


Рисунок 3.14 - Графічне відображення проекції поліноміальної регресії степеня 2

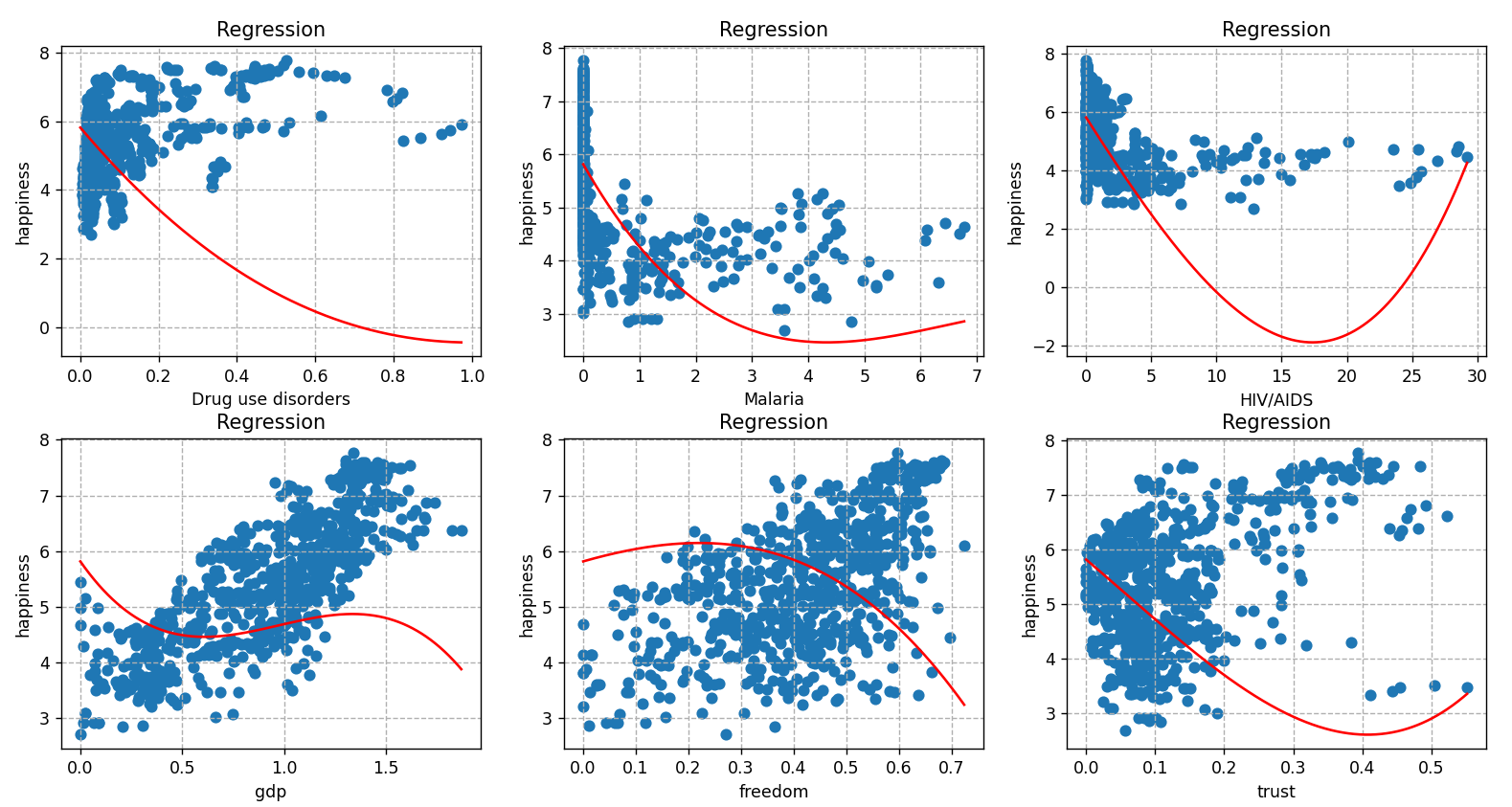


Рисунок 3.14 - Графічне відображення проекції поліноміальної регресії степеня 3

Отже, результатом роботи над цим розділом стала побудована модель регресійного аналізу, причому серед усіх протестованих вибрана найбільш точна. Досягнуто точності 80% за критерієм .

**ВИСНОВКИ**

В результаті виконання курсової роботи було розроблено алгоритм, який допомагає виявити чи новина фекова. Розглянуто основні підходи для реалізація такоє моделі. Для реалізація поставленої задачі було використано мову програмування Python та різні бібліотеки: pandas, numpy, matlplot, seaborn, nltk, gensim, tensorflow.keras та інші.

На основі детального опису та проведеного аналізу предметної області інтелектуального аналізу даних для визначення фейкових новин було отримано результати з високою точністю передбачення таких новин. Підтвердженням даних висновків є результати точності виявлення фейкових новин, яка складає 0.9939866369710467 для тестового набору даних. Результати досліджень показують, що дана можель може застосовуватися за призначенням і приносити користь.

Отже, поставлені задачі були виконані, а також планується створення датасету з російськими новинами та фековими статтями, які можна буде використовувати для тренування розробленої моделі. Дана можель може допомогти розробити телеграм-бот, який допоможе фільтрувати новині стрічки.

## 

## ПЕРЕЛІК ПОСИЛАНЬ

1. <https://stackoverflow.com/questions/51956000/what-does-keras-tokenizer-method-exactly-do>
2. <https://machinelearningknowledge.ai/keras-tokenizer-tutorial-with-examples-for-fit_on_texts-texts_to_sequences-texts_to_matrix-sequences_to_matrix/#>
3. <https://keras.io/api/preprocessing/>
4. <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>
5. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
6. Лекційні матеріали
7. Machine Learning. Coursera. Author: Andrew Ng
8. <https://keras.io/api/layers/activations/>
9. <https://www.tensorflow.org/guide/keras/sequential_model>
10. <https://en.wikipedia.org/wiki/Vanishing_gradient_problem>
11. <https://en.wikipedia.org/wiki/Backpropagation>

## ДОДАТОК А ТЕКСТИ ПРОГРАМНОГО КОДУ

<https://github.com/Dichik/FakeNewsML> - public repository with resources and code.