МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

імені Тараса Шевченка

Економічний факультет

Кафедра Статистики, інформаційно-аналітичних систем і демографії

ЗВІТ З ВИРОБНИЧОЇ ПРАКТИКИ

ОКР «Бакалавр»

ОП «Економічна аналітика та статистика»

студента: Зеленчука Андрія Андрійовича

База практики Державна служба статистики України

Керівник від бази практики	Керівник від кафедри
Директор департаменту статистики сільського господарства та навколишнього середовища	Кандидат економічних наук, доцент
(Посада)	(Посада)
Прокопенко Олег Миколайович	Моторина Тетяна Михайлівна
(П.І.Б.)	(П.І.Б.)
_____	_____
(Підпис)	(Підпис)

КИЇВ 2023

ЗМІСТ

ВСТУП

Метою виробничої практики студента освітньої програми «Економічна аналітика та статистика» економічного факультету Київського національного університету імені Тараса Шевченка є узагальнення раніше набутих теоретичних знань та вмінь, отримання нових практичних навичок професійної діяльності завдяки індивідуальній розрахунково-аналітичній роботі.

Основними завданнями практики студента є:

- вивчення функціональної структури Державної служби статистики України;
- ознайомлення з планом статистичних робіт департаменту (відділу), структурою та функціональними обов'язками, а також регулярними публікаціями (статистичні збірники, експрес-випуски, бюлетені);
- збір, обробка та аналіз статистичної інформації для виконання індивідуального розрахунково-аналітичного завдання;
- використання складних методів аналізу даних у індивідуальному розрахунково-аналітичному завданні.

Таким чином студент, окрім отримання знань та навичок, може набути і досвіду роботи, а також отримати розуміння того що саме відбувається у компаніях, зокрема у Державній службі статистики України, що є важливим та потрібним для них на даний момент, а також розширення кола знайомих, об'єднаними спільними професійними інтересами, аби в майбутньому за допомогою них можна було максимально швидко й ефективно вирішувати певні питання.

1. ОРГАНІЗАЦІЙНА СТРУКТУРА ДЕРЖАВНОЇ СЛУЖБИ СТАТИСТИКИ УКРАЇНИ

Згідно з Положенням про Державну службу статистики України Держстат є спеціально уповноваженим центральним органом виконавчої влади у сфері статистики, діяльність якого спрямовується та координується Кабінетом Міністрів України та який бере участь у формуванні державної політики у сфері статистики і забезпечує її реалізацію. Координацію роботи Держстату забезпечує Міністр Кабінету Міністрів України. [1]

Основними завданнями Держстату є:

1) участь у формуванні державної політики у сфері статистики та забезпечення її реалізації;

2) гармонізація системи державної статистики з міжнародними та європейськими нормами і стандартами;

3) збирання, збереження, оброблення, аналіз, захист та поширення офіційної державної статистичної інформації щодо масових явищ і процесів, які відбуваються в економічній, соціальній, демографічній, екологічній, культурній та інших сферах життя суспільства в Україні та її регіонах;

4) забезпечення актуальності, точності і надійності, узгодженості і порівнянності, доступності і ясності офіційної державної статистичної інформації;

5) розроблення, вдосконалення і впровадження статистичної методології, оприлюднення метаданих;

6) забезпечення розроблення, запровадження та вдосконалення системи статистичних класифікацій, які використовуються для проведення статистичних спостережень;

7) впровадження новітніх інформаційних технологій у процеси виробництва та поширення офіційної державної статистичної інформації;

8) взаємодія інформаційної системи органів державної статистики з інформаційними системами державних органів, органів місцевого самоврядування, інших юридичних осіб, міжнародних організацій та статистичних служб іноземних держав шляхом взаємного обміну інформацією, проведення методологічних,

програмно-технологічних та інших робіт, спрямованих на ефективне використання інформаційних ресурсів;

9) координація дій державних органів, органів місцевого самоврядування та інших юридичних осіб щодо організації діяльності, пов'язаної із збиранням та використанням адміністративних даних;

10) забезпечення принципу статистичної конфіденційності під час виробництва та поширення офіційної державної статистичної інформації. [1]

Держстат очолює Голова, який призначається на посаду та звільняється з посади Кабінетом Міністрів України. Голова Держстату має заступників, які призначаються на посаду та звільняються з посади Кабінетом Міністрів України відповідно до законодавства про державну службу.

Для підготовки рекомендацій щодо виконання завдань Держстату може утворюватися колегія. Рішення колегії можуть бути реалізовані шляхом видання відповідного наказу Держстату.

Для розгляду наукових рекомендацій та проведення фахових консультацій з основних питань діяльності в Держстаті можуть утворюватися інші постійні або тимчасові консультативні, дорадчі та інші допоміжні органи.

Рішення про утворення чи ліквідацію колегії, інших постійних або тимчасових консультативних, дорадчих та інших допоміжних органів, їх кількісний та персональний склад, положення про них затверджує Голова Держстату.

Правовою основою державної статистичної діяльності є закони України про інформацію та про офіційну статистику, згідно яких завдання державної статистичної діяльності полягає у своєчасному забезпеченні неупередженою та об'єктивною офіційною державною статистичною інформацією, необхідною для інформування суспільства, формування і моніторингу економічної та соціальної політики, прийняття обґрунтованих рішень державними органами на підставі результатів державних статистичних спостережень в інтересах забезпечення сталого розвитку, економічного добробуту та прав людини, виконання Україною взятих на себе зобов'язань в рамках чинних міжнародних угод, а також здійснення наукових досліджень. [2] [3]

2. ДІЯЛЬНІСТЬ ДЕПАРТАМЕНТУ СІЛЬСЬКОГО ГОСПОДАРСТВА ТА НАВКОЛИШНЬОГО СЕРЕДОВИЩА

Виробнича практика проходила в Департаменті статистики сільського господарства та навколишнього середовища апарату Державної служби статистики України. Директором департаменту є Прокопенко Олег Миколайович. У департаменті створені та функціонують такі відділи:

- відділ економічних рахунків сільського господарства;
- відділ статистики виробництва продукції сільського господарства;
- відділ статистики використання сільськогосподарської продукції;
- відділ структурних обстежень у сільському господарстві;
- відділ екологічних рахунків і статистики навколишнього середовища.

Згідно з Положенням про департамент статистики сільського господарства та навколишнього середовища, Департамент статистики сільського господарства та навколишнього середовища є самостійним структурним підрозділом апарату Державної служби статистики України, що забезпечує виконання завдань з організації та методології проведення державних статистичних спостережень у сфері сільського, рибного, лісового, мисливського господарства та навколишнього природного середовища. Департамент безпосередньо підпорядковується Голові Держстату. [4]

Основні завдання департаменту:

- участь у реалізації державної політики у сфері статистики сільського, рибного, лісового, мисливського господарства та навколишнього природного середовища та підготовці пропозицій щодо її формування;
- розроблення, удосконалення та впровадження статистичної методології у сфері статистики сільського, рибного, лісового, мисливського господарства та навколишнього природного середовища;
- сприяння координації дій державних органів, органів місцевого самоврядування та інших юридичних осіб, територіальних органів Держстату в питаннях організації діяльності, пов’язаної зі збиранням,

опрацюванням, поширенням і використанням статистичної інформації та адміністративних даних у сфері статистики сільського, рибного, лісового, мисливського господарства та навколишнього природного середовища;

- опрацювання, аналіз, поширення, захист і використання статистичної інформації щодо масових економічних явищ і процесів у сільському, рибному, лісовому та мисливському господарстві, а також екологічних явищ і процесів, які відбуваються в Україні та її регіонах;
- забезпечення відповідності статистичної інформації критеріям якості, підготовка звітів для користувачів щодо якості статистичних даних у сфері сільського, рибного, лісового, мисливського господарства та навколишнього природного середовища;
- забезпечення доступності, гласності й відкритості статистичної інформації, методології її складання та джерел даних. [4]

Відповідно до основних завдань департамент:

- бере участь у розробленні та реалізації проєктів довгострокових програм розвитку державної статистики, плану державних статистичних спостережень та інших планів, що регламентують роботу Держстату та його самостійних структурних підрозділів, з питань, віднесених до повноважень департаменту;
- організовує та проводить державні статистичні спостереження за економічними процесами та щодо стану навколишнього природного середовища
- забезпечує в межах повноважень формування інформаційної бази для прогнозування й аналізу тенденцій і закономірностей розвитку сільського, рибного, лісового, мисливського господарства та стану навколишнього природного середовища;
- здійснює моніторинг потреб користувачів у статистичній інформації щодо стану сільського, рибного, лісового, мисливського господарства та навколишнього середовища тощо. [4]

З питань, віднесених до повноважень департаменту, респонденти подають статистичні звіти:

- про площі та валові збори сільськогосподарських культур, плодів, ягід і винограду;
- про використання добрив і пестицидів;
- про виробництво продукції тваринництва, кількість сільськогосподарських тварин і забезпеченість їх кормами;
- про основні економічні показники роботи сільськогосподарських підприємств;
- про об’єкти погосподарського обліку;
- про наявність сільськогосподарської техніки;
- про витрати на виробництво продукції (робіт, послуг) сільського господарства;
- про надходження сільськогосподарських тварин на переробні підприємства;
- про надходження молока сирого на переробні підприємства;
- про перероблення винограду та виноматеріали;
- про реалізацію продукції сільського господарства;
- про надходження культур зернових і зернобобових та олійних на перероблення та зберігання;
- про добування водних біоресурсів;
- про відтворення та захист лісів;
- про облік, добування та розведення мисливських тварин;
- про викиди забруднюючих речових і парникових газів в атмосферне повітря від стаціонарних джерел викидів;
- про витрати на охорону навколишнього природного середовища;
- про утворення та поводження з відходами.

На основі звітів здійснюється оцінка макроекономічних та аналітичних показників, формується статистична інформація, яка публікується на вебсайті Державної служби статистики України, надається користувачам на запити.

3. ІНДИВІДУАЛЬНА РОЗРАХУНКОВО-АНАЛІТИЧНА РОБОТА

3.1. Вступ

В індивідуальній розрахунково-аналітичній роботі передбачалося здійснити аналіз первинних даних кількості викидів забруднюючих речовин у повітря за 2020 рік з використанням програмного забезпечення PyCharm, мови програмування Python та її бібліотек.

Для забезпечення повної конфіденційності даних у аналізі не буде вказано жодних повних назв окремих суб'єктів, їх код території чи вид діяльності.

На жаль, зважаючи на реалії сьогодення, для проведення аналізу можливо було використати лише наявні первинні дані щодо кількості викидів забруднюючих речовин за 2020 рік, які не є конфіденційними відповідно до Закону України «Про офіційну статистику». І хоча ці дані не є найбільш актуальними на сьогодні, все ж на їх основі можна провести якісний аналіз первинних даних, розробити код, який можна буде використати для даних за інші роки, а також модель машинного навчання.

Загальні етапи створення моделі машинного навчання:

1. Збір даних. На цьому етапі обираються потрібні дані, а потім збираються за допомогою визначеного алгоритму або вручну.
2. Обробка даних. На цьому етапі дані попередньо обробляються шляхом обробки всіх пропущених значень, категоріальних даних тощо. Також на цьому етапі відбувається масштабування даних.
3. Побудова моделі. На цьому етапі обирається відповідний алгоритм для створення моделі та будується безпосередньо модель.
4. Оцінка моделі. Після створення моделі вона оцінюється за допомогою обраних метрик, які найкраще оцінюють дані та модель.
5. Збереження та подальше використання моделі. Після успішної оцінки моделі вона зберігається для тестування та використання.

3.2. Розгляд даних

Дані, які будуть використовуватися для проведення аналізу, отримані органами державної статистики від респондентів під час проведення державного статистичного спостереження щодо викидів забруднюючих речовин і парникових газів в атмосферне повітря за формою № 2-ТП (повітря) (річна) "Звіт про викиди забруднюючих речовин і парникових газів у атмосферне повітря від стаціонарних джерел викидів".

Для початку розглянемо дані, які містяться у файлах та які будуть використовуватись у подальшому аналізі:

- KDMO – id суб’єкта підприємницької діяльності (місцевої одиниці)
- TE – код території за Класифікатором об’єктів адміністративно-територіального устрою України фактичного місця здійснення діяльності суб’єкта підприємницької діяльності
- KDG – код Єдиного державного реєстру підприємств та організацій України
- NU – назва суб’єкта підприємницької діяльності
- K00000 – K19000 – кількість викидів певної забруднюючої речовини, у тоннах.
- SEK_VDF10 – секція виду економічної діяльності суб’єкта підприємницької діяльності відповідно до КВЕД-2010
- VDF10 – код (розділ та група) виду економічної діяльності суб’єкта підприємницької діяльності відповідно до КВЕД-2010.

Тобто, у файлах вказана загальна інформація щодо кожного суб’єкта підприємницької діяльності, що подавав звіт, та його викиди забруднюючих речовин за певною класифікацією, у тоннах.

Наступним етапом буде оброблення отриманих даних.

3.3. Обробка даних

В одному з отриманих файлів із даними наявні ідентифікатор суб'єкта, коди ЄДРПОУ та КОАТУУ, текстові дані, числові дані з трьома цифрами після коми, секція за класифікатором КВЕД-2010, а також за цим же класифікатором розділ та група. Оскільки розділ та група знаходяться в одному стовпчику, можна їх розділити для більш загального аналізу суб'єктів з одного розділу. Секцію, дані щодо якої надано лиш одною буквою, можна змінити на числові дані, оскільки ці дані є класифікаційними, однак для збереження розуміння даних їх можна залишити. Числові дані для деяких суб'єктів кількості викидів певних забруднюючих речовин є відсутніми. Важливо мати контекст, оскільки може бути декілька варіантів:

1. Суб'єкт не може ніяким чином викидати речовини такого типу, а тому дані і не вказувались, тож пропущеним значенням варто присвоїти значення None.
2. Суб'єкт за певних причин не вказав жодне значення, і фактично воно може бути не нульовим. За цього випадку можна відкинути або суб'єктів, або тип забруднюючих речовин, аби не мати відсутні дані. Менш рекомендованим, однак теж методом є присвоїти відсутнім даним нульові значення.

Окрім цього варто також визначити чи є викиди у наявних даних та масштабувати їх.

Значення наявних кодів також можна обробити, оскільки вони зазвичай мають свою систему створення, а отже, один код можна розділити на декілька, по яким можна і узагальнити дані при аналізі.

І останнім типом даних є текстовий, який досить важко обробити. Для початку поверхнево переглянемо ці дані. Вони мають повні назви суб'єктів, і у цих назвах в більшості випадків відсутні букви «і»: «ТОВАРИСТВО З ОБМЕЖЕНОЮ В_ДПОВ_ДАЛЬН_СТЮ», «СТРУКТУРНИЙ П_ДРОЗД_Л». Тобто це не є одиничною помилкою, а є проблемою з кодуванням, де певні букви відсутні. Якщо переглянути нижче, то це виглядає іншим чином: «ФЎЛЎЯ», «ДНЎПРОВСЬКО°», «ДЕРЖАВНЕ ПЎДПРИЇМСТВО». Різниця у замінниках українських літер швидше всього зв'язана з тим, що дані, які у файлі розташовані з самого верху, були додані в інший час, ніж інші, або іншою людиною, без єдиного позначення. В першу чергу ці

проблеми будуть вирішені; з процесів першочергової обробки тексту це все, що варто зробити.

В іншому отриманому файлі наявні дані, що характеризують ідентифікатор суб'єкта, та числові дані з трьома цифрами після коми. Для початку потрібно визначити наявність викидів та масштабувати дані, аби побудована модель була менш залежна від значень кількості викидів певної забруднюючої речовини.

Для подальших дій було використано мову програмування Python у програмному забезпеченні PyCharm Professional, яке є безкоштовним для студентів. Також є версія PyCharm Community, яка з точки зору новачків у програмуванні нічим не відрізняється від розширеної версії.

Також окремо варто встановити саму мову програмування, та при створенні проекту у PyCharm вказати шлях до директорії з установленим Python та налаштувати створення віртуального навколишнього середовища.

У новоствореному файлі спочатку вкажемо тип кодування:

```
# coding: utf-8
```

Першим ділом варто прочитати дані з файлів. Для цього використовується бібліотека OpenPyXL. [5]

На жаль, ця бібліотека не підтримує файли формату .dbf, хоч вони і є форматами пакету Microsoft Excel, а тому завдяки цьому програмному забезпеченню відкриваємо та зберігаємо файл у форматі .xlsx.

Тепер можна прочитати дані з файлів:

```
from openpyxl import load_workbook
# read data
workbook = load_workbook('file_1.xlsx')
worksheet = workbook['Sheet']
```

Ось все, що потрібно зробити, аби відкрити файл, і тепер ми можемо читати дані або їх модифікувати. Однак значно легше обробляти табличні дані завдяки бібліотеці pandas. Щоб не перетворювати тип змінної з одного формату на інший можна застосувати певну хитрість – відкрити файл за допомогою функції бібліотеки pandas, однак як «двигун» вказати бібліотеку OpenPyXL. [7]

```
import pandas as pd
# save data
```

```
    DF = pd.read_excel(io='file_1.xlsx', sheet_name='Sheet',
header=0, usecols='A:EM', engine='openpyxl')
    print(DF.columns)
    print(f'Number of shapes: {DF.shape}')
    print(f'Number of cells: {DF.size}')
```

Та маємо вивід:

```
    Index(['KDMO', 'TE', 'KDG', 'NU', 'K00000', 'K01000',
'K01001', 'K01002',
           'K01003', 'K01004',
           ...
           'K18001', 'K18002', 'K18003', 'K18004', 'K18005',
'K18006', 'K18007',
           'K19000', 'SEK_VDF10', 'VDF10'],
          dtype='object', length=143)
    Number of shapes: (9644, 143)
    Number of cells: 1379092
```

Отже, таблицю було успішно відкрито. Маємо 143 стовпчики, 9644 рядків даних, та в загальному більше одного мільйону клітинок з даними чи без. Далі розділимо стовпчик з кодом КВЕД-2010 на 2 стовпчики, розділ (DIV) та групу (GROUP):

```
    # split column into columns
    DF[['DIV', 'GROUP']] =
DF['VDF10'].astype(str).str.split('.', expand=True)
```

Замінимо літери на потрібні, оскільки початковий тип кодування невідомий, а тому не можна перекодувати дані.

```
    # change characters
    for old_char, new_char in [('_', 'I'), ('Ў', 'I'), ('Ї',
'Є'), ('°', 'Ї')]:
        DF['NU'] = DF['NU'].str.replace(old_char, new_char)
```

Правильний код ЄДРПОУ містить 8 цифр, однак в даних досить багато і неправильних, що містять 6 чи 7 цифр, а тому їх не зовсім коректно обробляти. Однак можна взяти інформацію з коду КОАТУУ, який кодує фактичне місце здійснення діяльності суб’єкта підприємницької діяльності. Перші два символи відображають місто або область України, і саме цю інформацію буде використано для узагальнення. Варто зазначити, що в числових даних у всіх програмах не відображається 0, якщо він йде першою цифрою у числі, а тому у нашому випадку код області буде мати або 1, або 2 цифри. Отже, створення нового стовпчика даних буде виглядати таким чином:

```
    DF['T_CODE'] = DF['TE'].astype(str).str[:-8]
```

Отже, залишаються числові дані щодо кількості викидів, та потрібно визначити що робити з пропущеними даними. Для початку переглянемо чи є види забруднюючих речовин, щодо яких усі дані або відсутні, або нульові.

```
for column in DF.columns:
    num_zero = (DF[column] == 0).sum()
    num_missing = DF[column].isna().sum()
    no_number = (num_zero + num_missing)/DF.shape[0]
    if no_number > 0.95:
        print(f'{column}: {no_number}')
```

Всього маємо 137 стовпчиків із числовими даними, серед яких 105 мають відсоток відсутніх значень або нульових значень більше 95%. Це пов'язано з детальним розподіленням суб'єктів щодо викидів ними забруднюючих речовин у повітря. Тому можна викинути лише ті стовпчики, у яких всі дані є такими, а таких у таблиці 17.

```
for column in DF.columns:
    num_zero = (DF[column] == 0).sum()
    num_missing = DF[column].isna().sum()
    no_number = (num_zero + num_missing)/DF.shape[0]
    if no_number == 1:
        DF.drop(columns=column)
```

Можливо, буде доречним відфільтрувати шкідливі речовини, оскільки залишається 120 стовпчиків, що не є комфортним, проаналізувати у письмовому вигляді, а також більшість із них мають низьку частку наявних даних. Для цього в окрему змінну збережемо назву стовпчика, кількість наявних змінних та суму цих змінних, а також відразу відсортуємо за кількістю наявних даних:

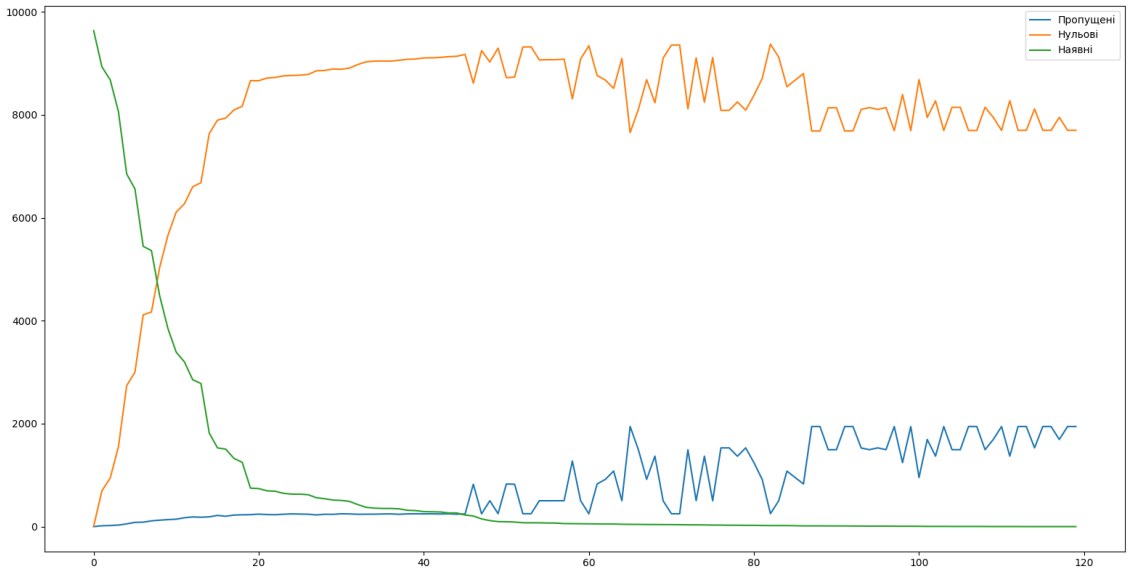
```
list_info = []
for column in DF.columns:
    if column not in ['KDMO', 'TE', 'KDG', 'NU',
'SEK_VDF10', 'VDF10', 'DIV', 'GROUP', 'T_CODE']:
        col_sum = DF[column].sum()
        col_digits = DF.shape[0] -
DF[column].isna().sum() - (DF[column] == 0).sum()
        list_info.append([column, col_digits, col_sum])
list_info.sort(key=lambda x: x[1], reverse=True)
for values in list_info[:10]:  # [:10] or without
    print(f'{values[0]}: {values[1]} in {values[2]}')
```

Тепер можемо візуалізувати дані. [8]

```
# graph zero/missing/value
```

```
    num_misses, num_zeros, num_values = [], [], []
    for column, count_data, sum_data in list_info:
        num_misses.append(DF[column].isna().sum())
        num_zeros.append((DF [column] == 0).sum())
        num_values.append(DF.shape[0] - DF
[column].isna().sum() - (DF [column] == 0).sum())
```

```
    plt.plot(num_misses, label='Пропущені')
    plt.plot(num_zeros, label='Нульові')
    plt.plot(num_values, label='Наявні')
    plt.legend()
    plt.show()
```

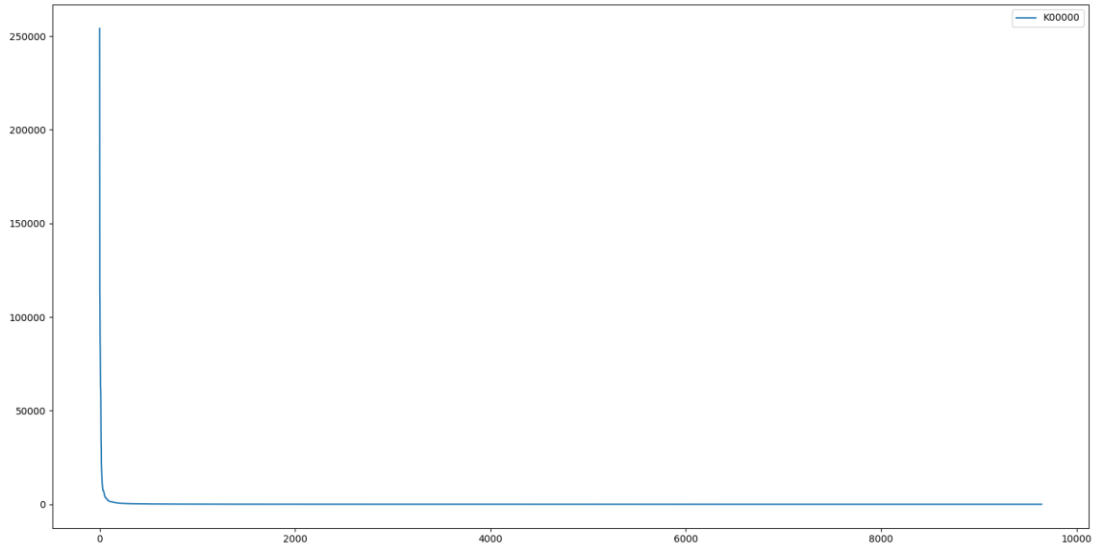


Як бачимо, лише близько щодо 10 видів шкідливих речовин є дані щодо викидів у половини суб'єктів підприємницької діяльності; у більшості ця кількість є дуже наближеною до нуля. Звісно, більшою мірою це спричинено саме нульовими значеннями, а не відсутніми, а нульові значення вказують на факт, що викидів у повітря даної речовини від певного суб'єкту у 2020 році не було, однак з точки зору аналітика вони є менш привабливими, ніж шкідливі речовини, щодо яких є більше наявних даних. Отже, для текстового опису візьмемо лише 10 видів шкідливих речовин, а усі дані застосуємо для певних моделей.

3.4. Аналіз даних

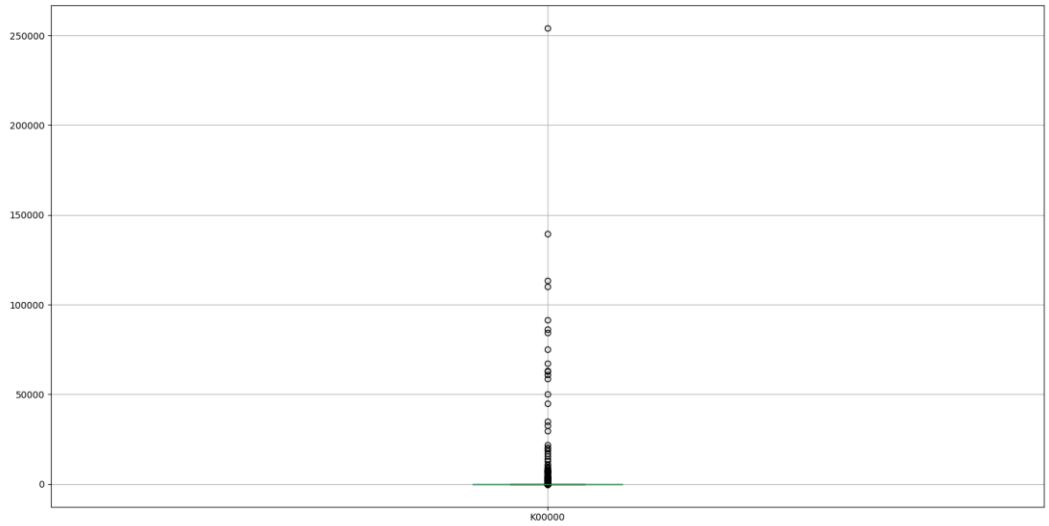
Отже, перейдемо до аналізу даних. Спробуємо побудувати дані на графіку. [7]

```
for column, count_data, sum_data in list_info[:10]:
    y_pos = DF.sort_values(column, ascending=False)
    y_pos.plot(y=column, use_index=False)
    plt.show()
```



Усі графіки є такого вигляду. Це не помилка у коді, це просто ось такі наявні дані. Для прикладу також побудуємо коробку з вусами:

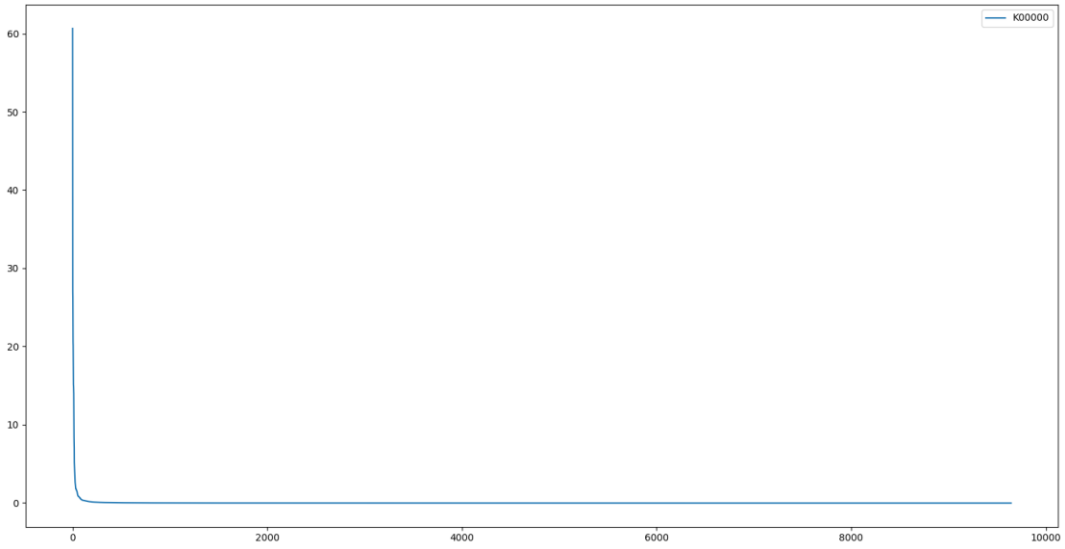
```
DF.boxplot(column)
```



Як бачимо, у даних дуже багато викидів. Але це одночасно і не викиди, адже такі компанії є, і вони дійсно стільки роблять викидів у повітря. Спробуємо зробити стандартизацію даних, а саме z-нормалізацію. Для цього застосуємо уже готову функцію fit_transform() з ініціалізованого StandardScaler() з пакету scikit-learn. [10]

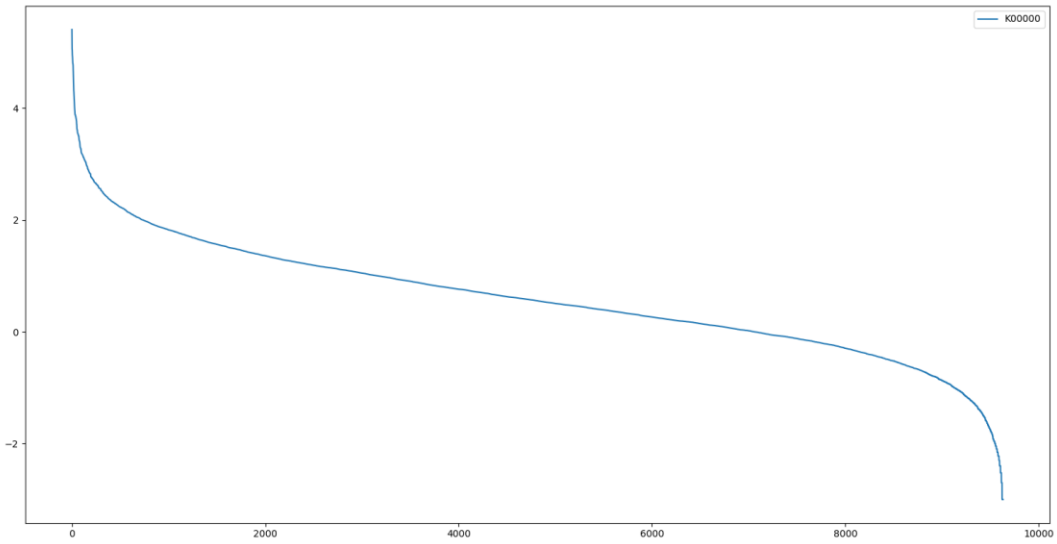
```
from sklearn.preprocessing import StandardScaler
```

```
    scaler = StandardScaler()
    for column, count_data, sum_data in list_info[:10]:
        DF[column] = scaler.fit_transform(DF[column][:,
np.newaxis])
        y_pos = DF.sort_values(column, ascending=False)
        y_pos.plot(y=column, use_index=False)
        plt.show()
```



Абсолютні значення даних значно зменшились, деякі дані стали від’ємними, оскільки за умовами даної нормалізації в результаті дані мають мати середнє значення 0, а стандартне відхилення 1. Це на графіку важко помітити, оскільки відношення між найбільшими значеннями та іншими даними досі залишається дуже великим. Тому замінимо її на логарифмічну. [6]

```
for column, count_data, sum_data in list_info[:10]:
    DF[column] = np.log10(DF[column])
    y_pos = DF.sort_values(column, ascending=False)
    y_pos.plot(y=column, use_index=False)
    plt.show()
```



І це вже виглядає більш адекватно.

Даний графік показує не саме значення даних, а кількість нулів у даних. Варто пам’ятати, що чим більше дані відрізняються від 1, тим більше вони спотворені, оскільки дані змінені не в якомусь звичайному відношенні, а в логарифмічному.

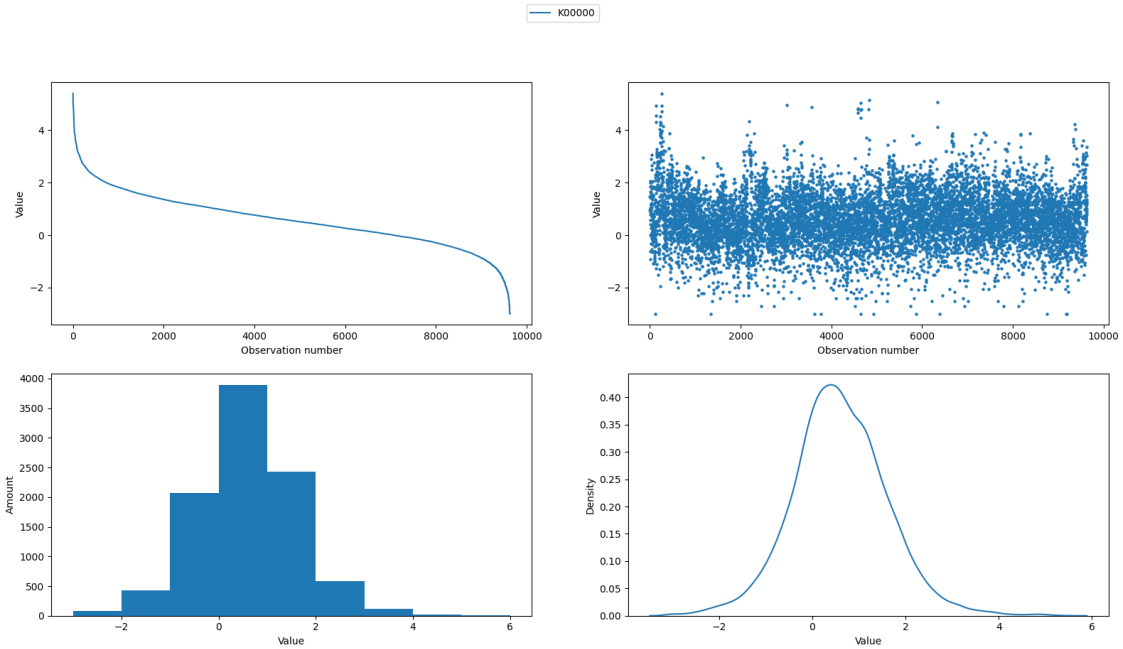
Однак саме такі дані можливо якось аналізувати, тож їх і побудуємо у рисунках.

Для початку відберемо лише наявні дані, тобто більше 0. Розрахуємо логарифм, відсортуємо, додамо колонку з номером відсортованих даних для одного з рисунків. Визначимо максимальне та мінімальне значення та на їх основі створимо змінну з кордонами стовпчиків для гістограми. Ініціалізуємо загальний рисунок з чотирма рисунками два на два та побудуємо їх, підписавши системи координат. [8] [9]

```
    for column, count_data, sum_data in list_info[:10]:
        DF = DF[DF[column] > 0]
        DF[column] = np.log10(DF[column])
        col_sorted = DF.sort_values(column, ascending=False)
        col_sorted['x'] =
pd.DataFrame(range(len(col_sorted[column])))
        fig, axes = plt.subplots(2, 2)
        col_sorted.plot(y=column, use_index=False, ax=axes[0,
0], legend=False)
        axes[0, 0].set(xlabel="Observation number",
ylabel="Value")
        data_min = np.floor(col_sorted[column].min())
        data_max = np.ceil(col_sorted[column].max())
        # Compute the bin edges based on the minimum and
maximum values
        bin_edges = np.arange(data_min, data_max + 1, 1)
        axes[1, 0].hist(col_sorted[column], bins=bin_edges)
        axes[1, 0].set(xlabel="Value", ylabel="Amount")
```

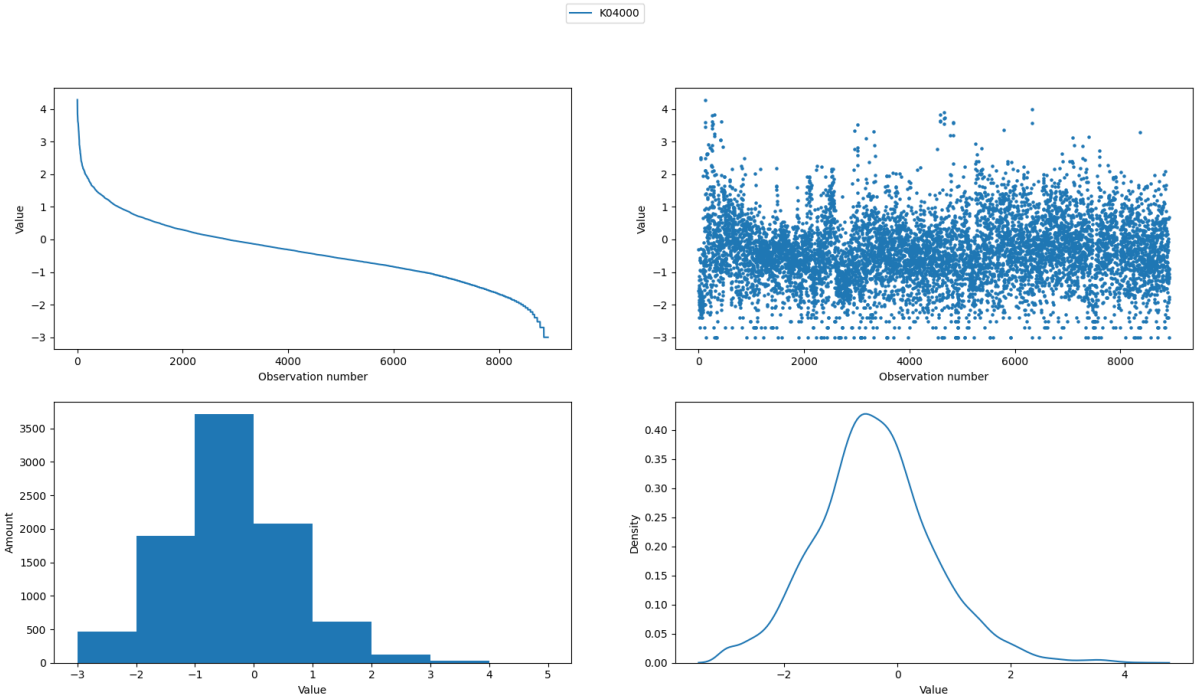
```
        axes[0, 1].scatter(col_sorted['x'],
col_sorted[column], s=5)
        axes[0, 1].set(xlabel="Observation number",
ylabel="Value")
        sns.kdeplot(col_sorted[column], ax=axes[1, 1])
        axes[1, 1].set(xlabel="Value", ylabel="Density")
        fig.legend(labels=[column], loc='upper center')
        plt.show()
```

Отримаємо для кожного з 10 обраних шкідливих речовин, що викидаються у повітря найбільшою кількістю суб'єктів підприємницької діяльності, по чотири рисунки.



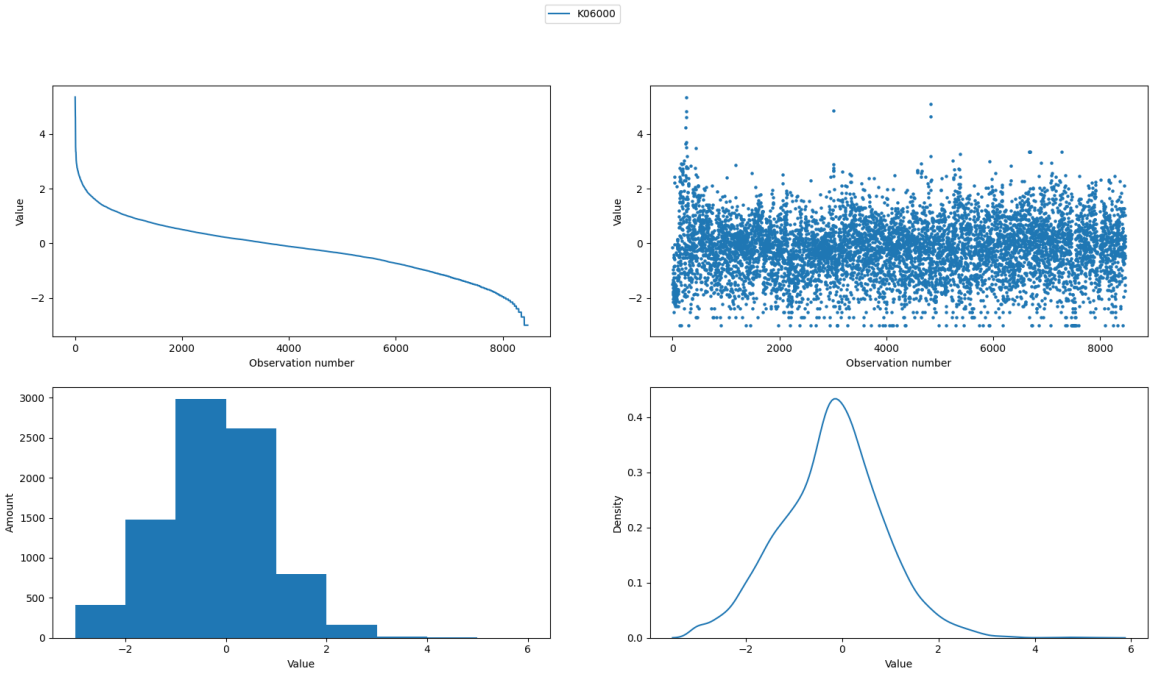
У верхньому лівому рисунку маємо сортовані значення викидів для кожного суб'єкта підприємницької діяльності, а зверху справа – не сортовані, щоб бачити в загальному як виглядають дані. Знизу зліва знаходиться гістограма, де ширина кожного стовпчика становить 1. Знизу справа знаходиться рисунок щільності.

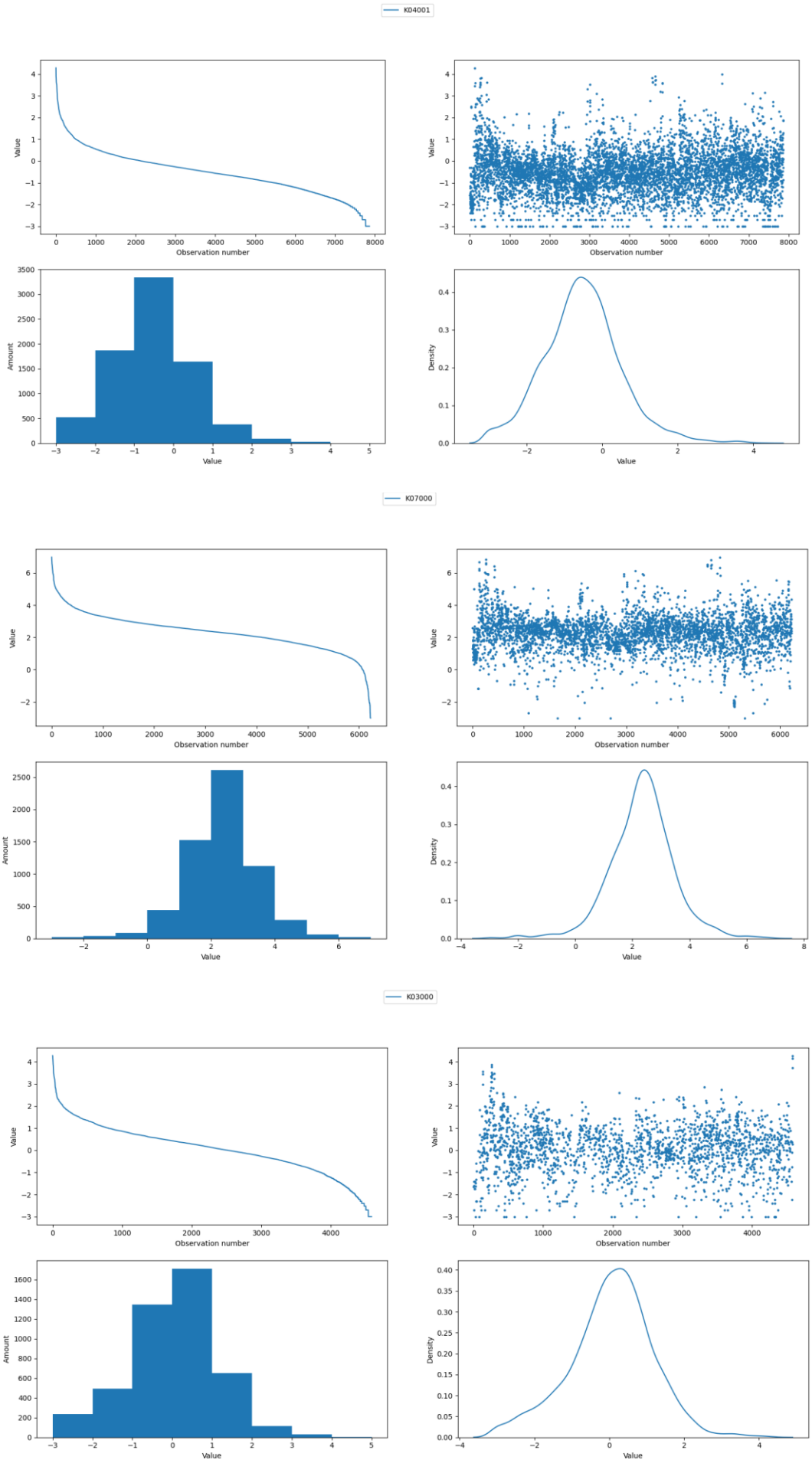
Отже, перейдемо до розгляду. Дані викидів речовини K00000 мають нормальний розподіл, і найбільша концентрація значень саме в діапазоні від 1 до 10 тонн.

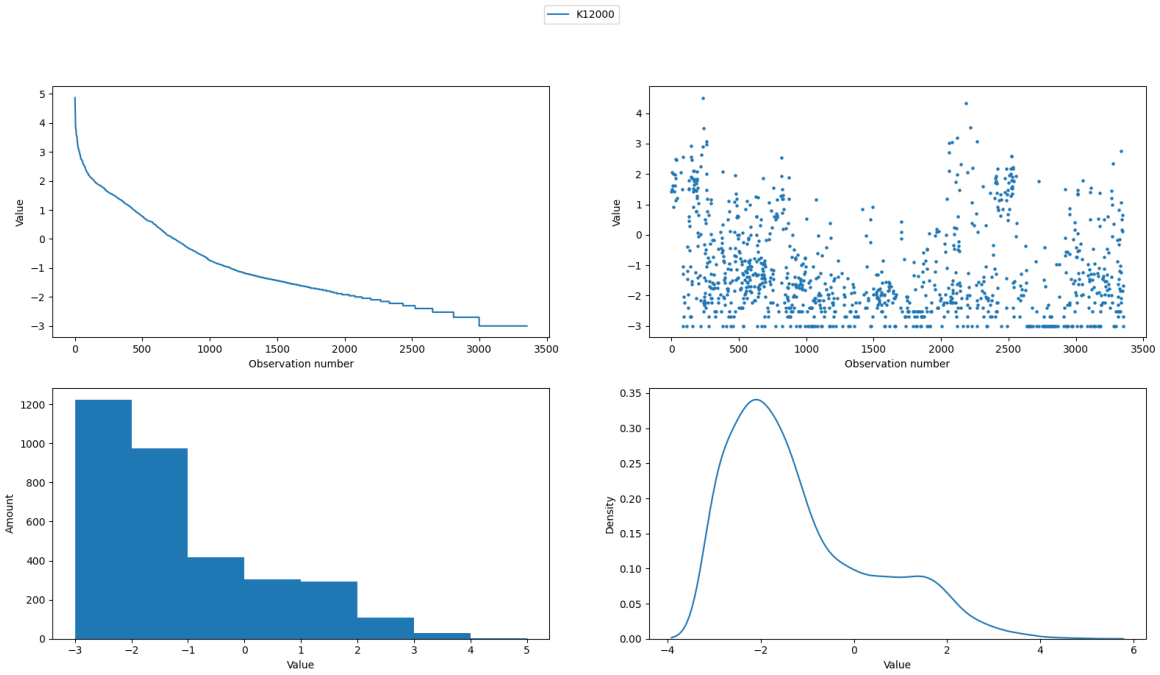


У речовини К04000 викиди також мають практично ідеальний нормальний розподіл, гостровершинний, з найбільшою концентрацію значень від 0,1 до 1 тонни.

Аналогічна ситуація і з наступними шкідливими речовинами, де змінюється лише кількість суб’єктів, що викидали, та обсяги.

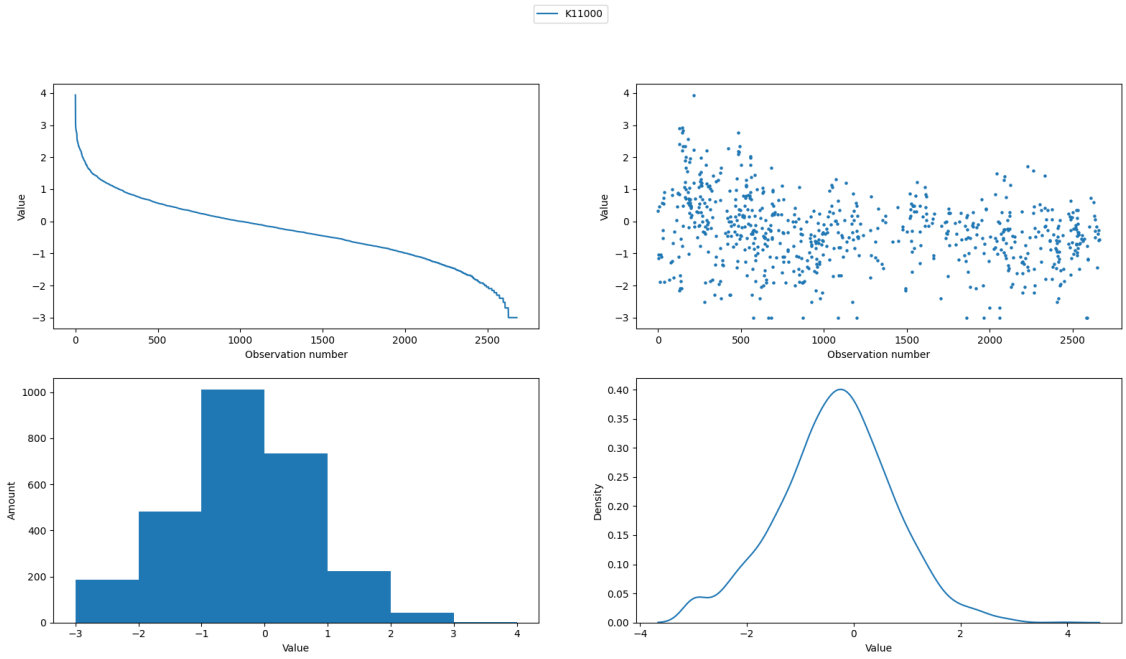


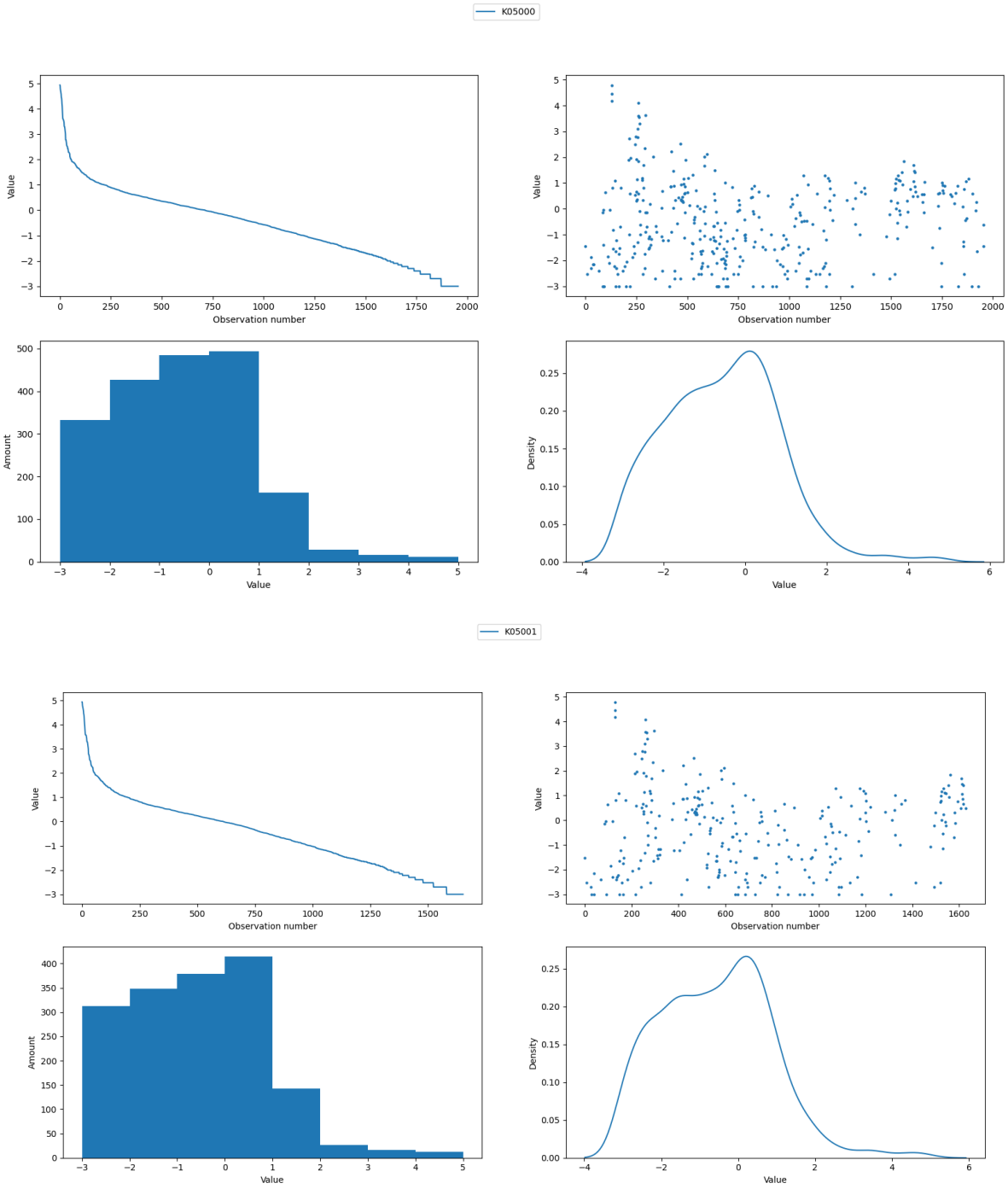

K04001
Value
Observation number
Amount
Density
K07000
K03000



Лиш нарешті в речовини К12000 розподіл не є нормальним, а як наче розподіл двовершинний. Зв’язано з тим, що найбільша концентрація у значень 0,001 – 0,01, а це ті значення, коли обсяги викидів хоч і є у багатьох, однак в дуже низьких обсягах.

Помітним є значне зменшення кількості суб’єктів, що викидають дану речовину порівняно з попередніми речовинами.





У двох останніх графіках графік щільності наближений до нормального, але має більш пологий схил зі сторони менших значень.

3.5. Побудова та оцінка моделі

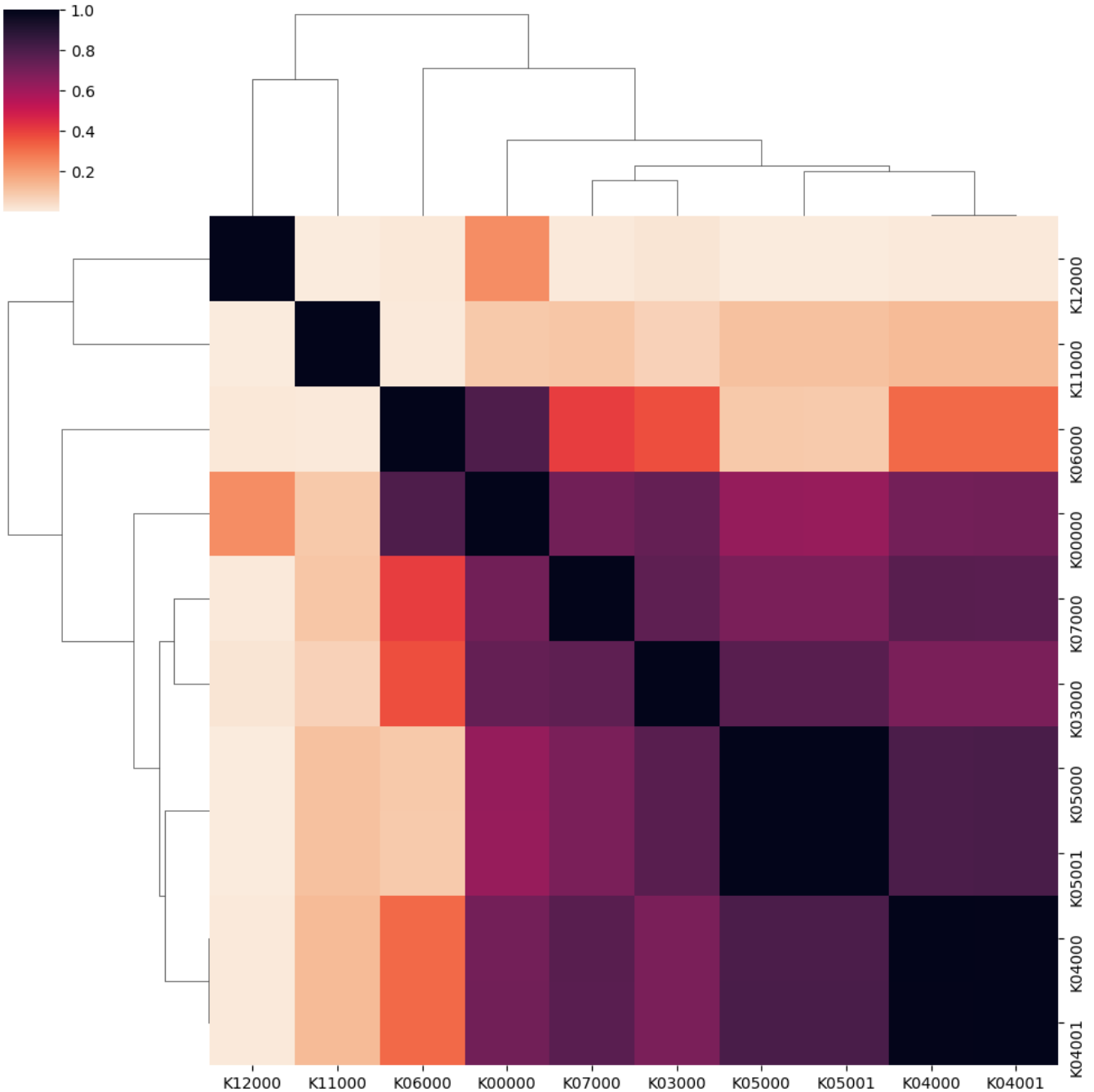
Отже, перейдемо до побудови моделі. Для початку варто розглянути з чим можна пов’язати дані:

- з даними щодо викидів, тобто, знайти кореляційний взаємозв’язок. Для цього можна не будувати модель, а створити кореляційну матрицю та проаналізувати її. Ті дані, в яких буде зв’язок, можна застосувати для побудування регресійної моделі;
- з регіоном, тобто визначити в якому регіоні, скільки та яка речовина викидається в повітря. Звісно, це буде теж лише для знаходження взаємозв’язків, оскільки не в усіх сферах діяльність компанії залежить від регіону та того, що знаходиться чи відбувається в його межах, починаючи від природних копалин у землі та завершуючи економічною ситуацією та поглядами політичної верхівки. А тому, якщо з’являється новий суб’єкт підприємницької діяльності, не можна на основі його викидів спрогнозувати з якого він регіону, чи навпаки;
- з секцією чи розділом основного виду економічної діяльності суб’єкта відповідно до класифікатора КВЕД-2010. А оскільки обсяги викидів більшою мірою і кількістю, і різновидом залежать від виду економічної діяльності, тут зв’язок може бути досить сильним, і при створенні нового суб’єкта економічної діяльності можна спрогнозувати його основний вид економічної діяльності чи навпаки, спрогнозувати викиди на основі економічної діяльності.

Порівняння даних викидів між собою та з регіоном, як результат надасть факт того, що з чим та як корелює, що можна застосувати як факт для інших роздумів. А порівняння даних викидів з секцією чи розділом КВЕД-2010 може надати модель машинного навчання, класифікатор, який на основі даних про викиди буде визначати основний вид економічної діяльності. Однак яка економічна цінність даної моделі? Її можна застосовувати для перевірки правдивості поданих обсягів викидів чи виду економічної діяльності суб’єкта, якщо припустити, що дані про викиди усі є правдивими.

Код, завдяки якому можна побудувати кореляційну матрицю до перших десяти речовин, що найбільше викидаються у повітря. Для матриці усі рядки з відсутніми значеннями відкидаються, тому у ній зображено результат для 9454 рядків даних, а для візуалізації використовується бібліотека seaborn. [9]

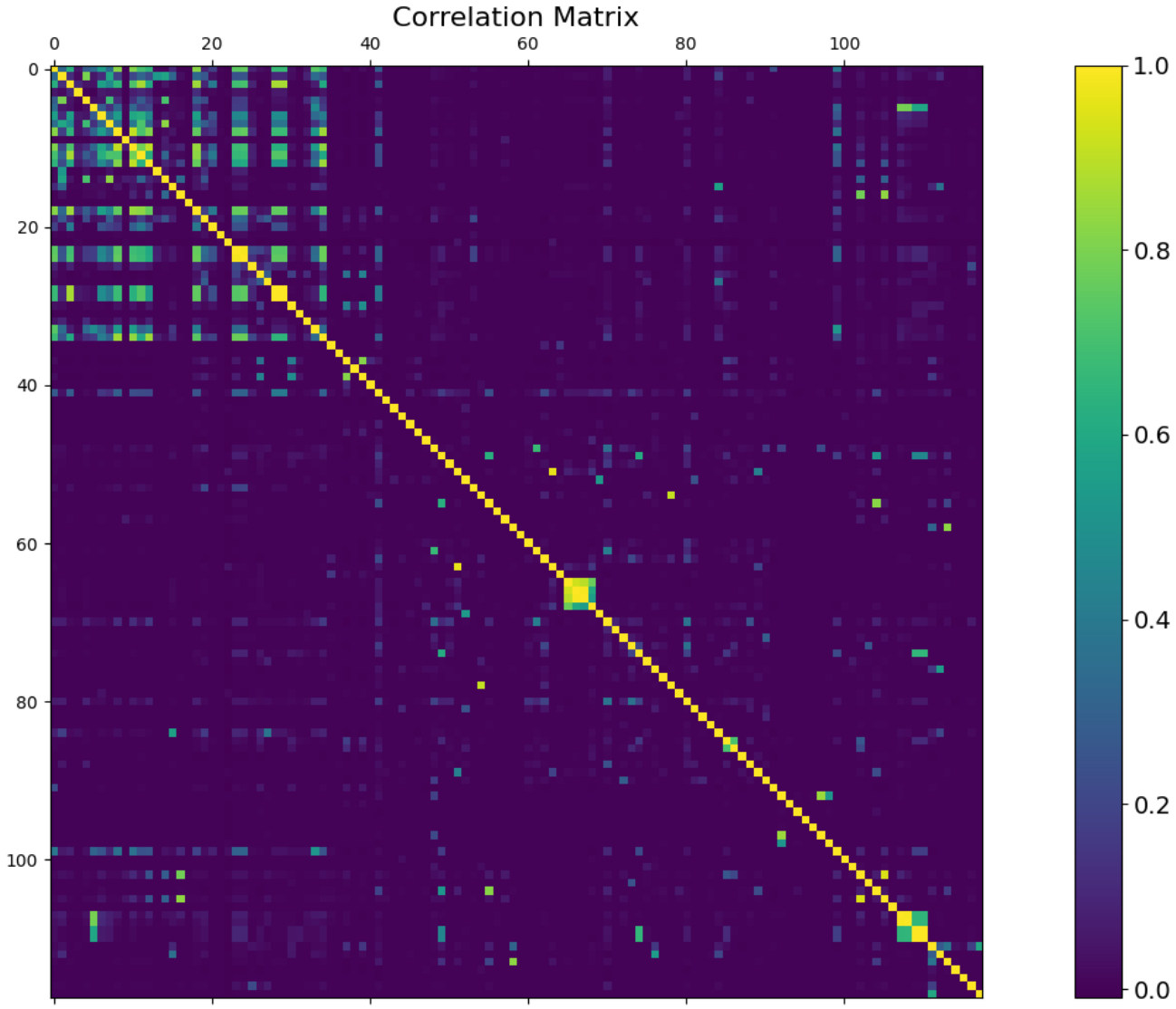
```
    DF_sorted = pd.DataFrame()
    for column, count_data, sum_data in list_info[:10]:
        DF_sorted[column] = DF[column]
    print(DF_sorted.dropna(how='any'))
    sns_plot =
sns.clustermap(DF_sorted.dropna(how='any').corr(),
cmap="rocket_r")
    plt.show(
```



Як бачимо, максимальна кореляція міститься між речовинами K04001 та K04000, а також між K05001 та K05000. На основі рисунку можна припустити, що все ж серед даних є розподіли чи підвиди речовин.

Створимо також кореляційну матрицю і для загальних даних. [8]

```
    DF_sorted = worksheet.drop(columns=['KDMO', 'TE', 'KDG',
'NU', 'SEK_VDF10', 'VDF10', 'GROUP', 'T_CODE', 'K11043',
'K13005'])
    f = plt.figure(figsize=(22, 18))
    plt.matshow(DF_sorted.dropna(how='any').corr(),
fignum=f.number)
    cb = plt.colorbar()
    cb.ax.tick_params(labelsize=14)
    plt.title('Correlation Matrix', fontsize=16)
    plt.show()
```



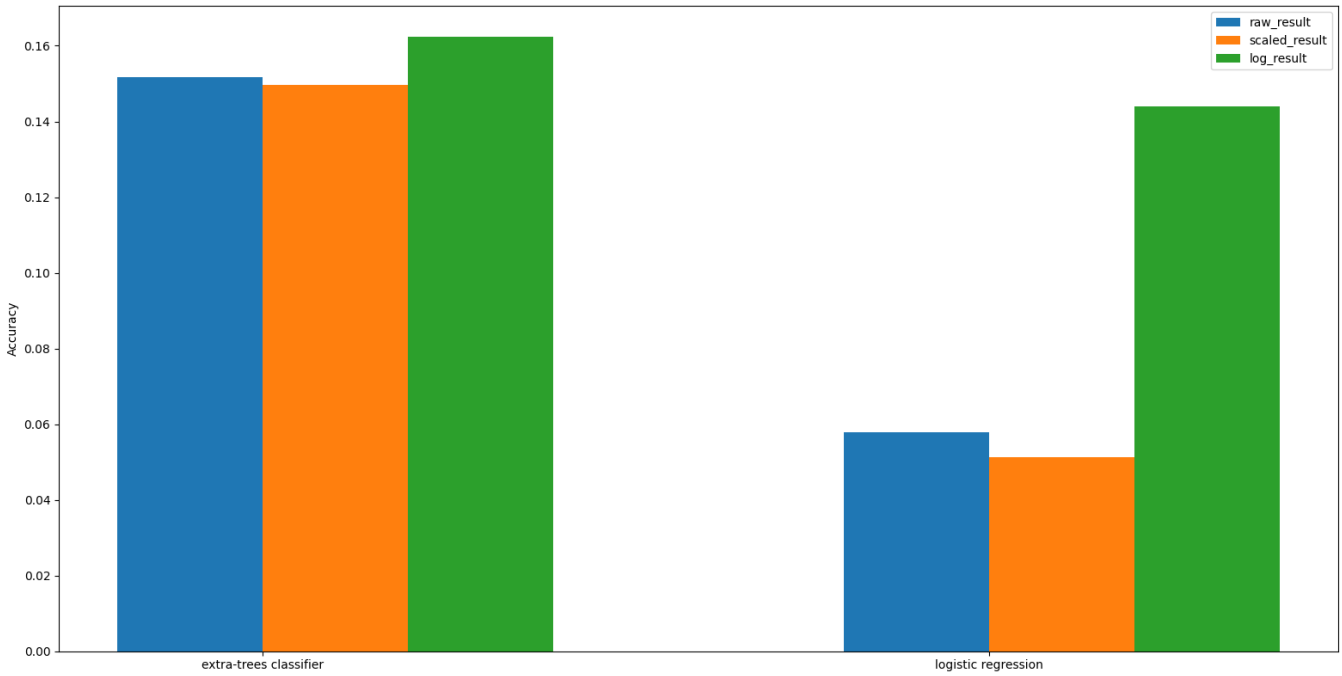
Тут вже залишилось 7696 рядків даних. Як бачимо, у досить невеликої кількості шкідливих речових є висока кореляція, і зазвичай це відбувається між сусідніми речовинами по назві, але кореляція є, попри те, що тут використовувались необроблені дані.

Тепер можна перейти до створення моделі, а саме мультикласифікатора розділу економічної діяльності суб'єкта за КВЕД-2010, маючи на основі значення щодо викидів шкідливих речових у тоннах. Візьмемо для порівняння і необроблені дані, і стандартизовані, і логарифмовані, а також для порівняння використаємо 2 різних способи тренування мультикласифікатора завдяки бібліотеці sklearn. [10]

```
    from sklearn.linear_model import LogisticRegression
    from sklearn.ensemble import ExtraTreesClassifier
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import classification_report,
balanced_accuracy_score

    scaler = StandardScaler()
    X = DF.drop(columns=['KDMO', 'TE', 'KDG', 'NU',
'SEK_VDF10', 'VDF10', 'GROUP', 'T_CODE', 'K11043',
'K13005'])
    X.dropna(how='any', inplace=True)
    y = X['DIV']
    X.drop(columns=['DIV'], inplace=True)

    X_scaled, X_log = pd.DataFrame(), pd.DataFrame()
    for column in X.columns:
        X_scaled[column] =
pd.DataFrame(scaler.fit_transform(X[column][:, np.newaxis]))
        X_log[column] = np.log10(X[column]+0.00001)
    X_scaled = X_scaled.copy()
    X_log = X_log.copy()

    raw_result, scaled_result, log_result = [], [], []
    for X, y, data_type in [(X, y, raw_result), (X_scaled, y,
scaled_result), (X_log, y, log_result)]:
        X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.3, random_state=42)
        ETC = ExtraTreesClassifier().fit(X_train, y_train)
        ETC_pred = ETC.predict(X_test)
        print(classification_report(y_test, ETC_pred))
```

```
        data_type.append(balanced_accuracy_score(y_test,
ETC_pred))
        LR = LogisticRegression(solver='lbfgs',
multi_class='multinomial', max_iter=70000).fit(X_train,
y_train)
        LR_pred = LR.predict(X_test)
        print(classification_report(y_test, LR_pred))
        data_type.append(balanced_accuracy_score(y_test,
LR_pred))

    x_labels = ['extra-trees classifier', 'logistic
regression']
    X_axis = np.arange(len(x_labels))
    plt.bar(X_axis - 0.1, raw_result, 0.2,
label='raw_result')
    plt.bar(X_axis + 0.1, scaled_result, 0.2,
label='scaled_result')
    plt.bar(X_axis + 0.3, log_result, 0.2,
label='log_result')
    plt.xticks(X_axis, x_labels)
    plt.xlabel("Model type")
    plt.ylabel("Accuracy")
    plt.legend()
    plt.show()
```

Завдяки функції classification_report() можна отримати метрики результатів щодо кожного класу, однак вони є складними для аналізу, тож використаємо загальну точність кожної з моделей.



Як бачимо, в загальному класифікатор додаткових дерев має вищу точність, ніж логістична регресія. Якщо зрівнювати по даним, то стандартизовані дані мають трохи гіршу точність, хоча в рази пришвидшують навчання моделі, оскільки логістична регресія на первинних даних вчиться дуже довго та їй не є достатнім менше 70 тисяч ітерацій. Модель, навчена на логарифмованих даних має вищу точність, і це помітно саме у логістичній регресії у зв’язку з більшою залежністю від викидів. Це також і компенсує зміну відношення даних між собою.

Однак все ще точність є дуже малою для використання цих моделей, а тому потрібно продовжувати покращувати їх.

ВИСНОВКИ

Отже, на основі первинних даних щодо викидів шкідливих речовин у повітря суб'єктами підприємницької діяльності у 2020 році було вивчено:

- як обробляти та візуалізувати дані за допомогою мови програмування Python та бібліотек NumPy, Pandas, Matplotlib, Seaborn;
- як будувати кореляційні матриці;
- що робити з даними, в яких є відсутні та нульові значення, а також викиди;
- як навчати моделі класифікації за допомогою бібліотеки sklearn;
- як впливають на класифікаційну модель метод навчання та спосіб обробки даних.

В загальному щодо результатів маємо, що:

- для кожної речовини окремо чим більше є наявних даних, тим більше її розподіл схожий на нормальний і не має різних схилів;
- між собою дані викиди речовин інколи дуже добре корелюють, навіть з різних груп;
- є потенціал до хорошого зв'язку між викидами суб'єкта та його розділом економічної діяльності за КВЕД-2010, оскільки для деяких класів окремо точність становить більше 80%.

На жаль, отримані моделі мають низьку точність, однак для покращення результатів можна спробувати:

- інше відношення розподілу даних на тренування та тестування, інше значення random_state при випадковому розподілі даних;
- інший алгоритм оптимізації для логістичної регресії;
- інші види класифікаторів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Положення про Державну службу статистики України [Електронний ресурс]. – 2014. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/481-2014-п#Text.
2. Закон України "Про інформацію" [Електронний ресурс]. – 1992. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/2657-12#Text.
3. Закон України "Про офіційну статистику" [Електронний ресурс]. – 2022. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/2524-20#Text.
4. Основні функції структурних підрозділів [Електронний ресурс] – Режим доступу до ресурсу: https://ukrstat.gov.ua/telefon/ukr/spc_n.htm.
5. Документація OpenPyXL [Електронний ресурс] – Режим доступу до ресурсу: https://openpyxl.readthedocs.io/en/stable/usage.html.
6. Документація NumPy [Електронний ресурс] – Режим доступу до ресурсу: https://numpy.org.
7. Документація Pandas [Електронний ресурс] – Режим доступу до ресурсу: https://pandas.pydata.org.
8. Документація Matplotlib [Електронний ресурс] – Режим доступу до ресурсу: https://matplotlib.org.
9. Документація Seaborn [Електронний ресурс] – Режим доступу до ресурсу: https://seaborn.pydata.org.
10. Документація Scikit-learn [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/.