Висновок до лекції 4

SQLite – це реалізація SQL, яка добре працює з Python. Замість того, щоб використовувати метод роботи з клієнтським сервером, використовуються з'єднання, встановлені між Python та базою даних SQL, створюючи об’єкт підключення SQL. Після створення об’єкта з'єднання використовується метод створення об’єкта курсору. Об'єкт курсору має методи SQLite, доступні для виконання операцій SQL в базі даних. SQLite може працювати в інтерактивному режимі з командного рядка, мова Python також може взаємодіяти з SQLite через імпорт модулів.

Питання для закріплення

1. З яких трьох мов спеціального призначення складається SQL?
2. Наведіть основні команди мови SQL.
3. Яке призначення утиліти csvsql?
4. Як отримати доступ до баз даних із програм, написаних на Python?

Список рекомендованої літератури

1. IoT Fundamentals: Big Data & Analytics // Електронний ресурс. Режим доступу: https://www.netacad.com/courses/iot/big-data-analytics
2. sqlcsv // Електронний ресурс. Режим доступу: https://pypi.org/project/sqlcsv/
3. Programming with Databases – Python // Електронний ресурс. Режим доступу: https://swcarpentry.github.io/sql-novice-survey/10-prog/index.html

Лекція 5.
Процедура імпорту даних із файлів у Pandas.
Імпорт даних з мережі Інтернет.
Засоби для кореляційного аналізу в Pandas

План лекції

5.1. Статистичні підходи до аналітики великих даних.
5.2. Використання Pandas.
5.3. Імпорт даних з файлів.
5.4. Імпорт даних з мережі Інтернет.
5.5. Описова статистика в Pandas.
5.6. Засоби для кореляційного аналізу в Pandas.

5.1. Статистичні підходи до аналітики великих даних

У аналітиці великих даних використовуються різні статистичні підходи. Як відомо, описова статистика описує вибірку. Це корисно для розуміння вибіркових даних та для визначення їх якості. При роботі з великою кількістю даних, що надходять із багатьох джерел, може виникнути багато проблем. Іноді точки даних можуть бути пошкоджені, неповні або повністю відсутні. Описова статистика може допомогти визначити, яка частина даних у вибірці є корисною для аналізу та визначити критерії для вилучення даних, які є невідповідними або проблемними. Графіки описової статистики є корисним способом швидкого судження про вибірку. Наприклад, для аналізу може бути обраний зразок твітів. Деякі твіти у зразку містять лише символи, а інші твіти містять символи та зображення. Ми можемо аналізувати твіти, які містять зображення або твіти без зображень. Це дозволить визначити недійсні твіти на основі дуже простого критерію. Точки даних, які не відповідають основним критеріям, будуть вилучені з вибірки до того, як розпочнеться аналіз.

Методи машинного аналізу дуже часто використовуються в аналітиці великих даних:

- **Кластерний аналіз** – використовується для пошуку груп спостережень, схожих між собою.
- **Асоціація** – використовується для пошуку спільних зустрічей значень для різних змінних.
- **Регресія** – використовується для кількісної оцінки взаємозв'язку між варіаціями однієї або декількох змінних, якщо такі є.

У машинному навчанні комп'ютерне програмне забезпечення або надається, або отримує власний набір правил, які використовуються для проведення аналізу. Методи машинного навчання можуть вимагати великої потужності обробки і стали життєздатними лише при наявності паралельної обробки.

На рис. 5.1. показана таблиця, що складається з двох полів. Одне поле містить змінну, а інше складається із статистики, яка описує значення цієї змінної. У

цьому прикладі десять учнів взяли вікторину на десять балів. Коли викладач аналізує бали, створюється розподіл балів, як показано у другій таблиці. Це виражає кількість разів, коли бал відбувся в класі. Ймовірність балу виражається у співвідношенні частоти балів до загальної кількості балів.

Студент	**Вікторина (10 балів)**	**Оцінка**	**Частота оцінки**	**Ймовірність балу**
Студент 1	6	1	0	0
Студент 2	7	2	0	0
Студент 3	7	3	0	0
Студент 4	8	4	0	0
Студент 5	7	5	1	0.1
Студент 6	9	6	1	0.1
Студент 7	10	7	4	0.4
Студент 8	8	8	2	0.2
Студент 9	7	9	1	0.1
Студент 10	5	10	1	0.1

Рис. 5.1. Приклад таблиць з оцінками учнів за виконану вікторину [1]

Розподіл частот складається з усіх унікальних значень змінної та кількості разів, коли значення наступають у наборі даних. У розподілах ймовірностей замість частот використовується пропорція часу, коли значення виникає в даних. Гістограма може представляти розподіл набору даних. У випадку дискретної змінної кожному біну гістограми присвоюється певне значення. У випадку безперервної дії кожен контейнер пов'язаний з діапазоном значень. В обох випадках висота бункера представляє кількість разів, коли значення змінної приймає задане значення або потрапляє в діапазон відповідно.

Представлення гістограми розподілу даних може приймати будь-яку форму. У разі безперервної змінної форма також залежатиме від ширини бункерів, тобто від їх діапазону. Деякі фігури можна моделювати за допомогою чітко визначених функцій, які називаються функціями розподілу ймовірності.

Функції розподілу ймовірностей дозволяють представляти форму всього розподілу набору даних, використовуючи лише невеликий набір параметрів, таких як середнє значення та дисперсія.

Функція розподілу ймовірностей, яка особливо підходить для відображення багатьох подій, що відбуваються в природі, – це Гауссовий, або нормальний розподіл, яке є симетричним і має форму дзвоника. Інші розподіли не симетричні. Вершина графіка може бути ліворуч або праворуч від центру. Ця властивість розподілу називається перекосом. Деякі розподіли матимуть два піки і відомі як бімодальні. Правий і лівий кінці графіка розподілу відомі як хвости. Однією з характеристик розподілів, яка дуже часто використовується, є міри центральної тенденції. Ці міри виражають значення, які має змінна, яка є найближчою до центральної позиції в розподілі даних. Загальні міри центральності – це середнє, медіана та мода. Мода вибірки даних – це значення, яке зустрічається найчастіше (рис. 5.2). Значення, які ближчі до центру розподілу, зустрічаються з більшою частотою.

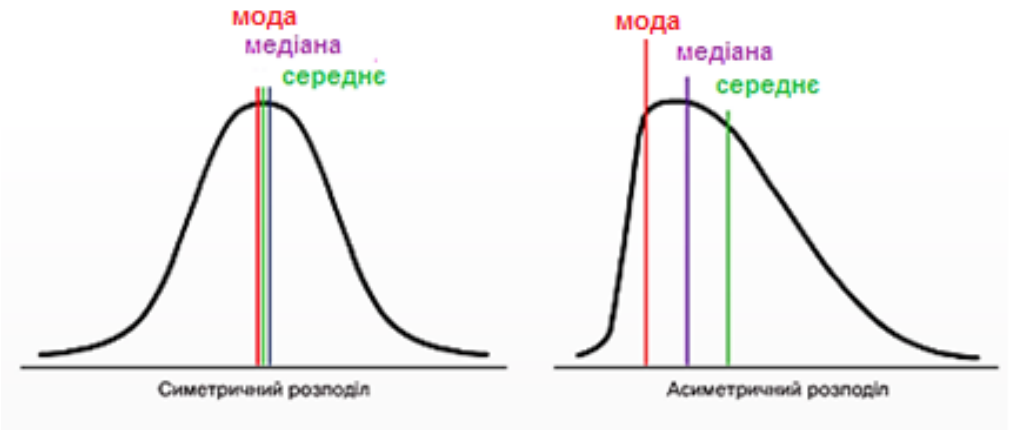


Рис. 5.2. Загальні міри центральності – це середнє, медіана та мода [1]

Середнє значення, також відоме як середнє, є найбільш відомим показником центральної тенденції. Дорівнює сумі всіх значень даних, поділених на кількість значень у наборі даних. Хоча середнє дуже часто використовується у повсякденному житті, вона, як правило, не є найкращим показником найбільш репрезентативного значення для розподілу. Наприклад, якщо в розподілі є незвично високі або низькі значення, на середнє можуть сильно впливають ті

крайні значення, які називаються викидами. Залежно від кількості випадаючих в наборі даних, середня величина або середня величина "перекошується" або змінюється в ту чи іншу сторону.

Медіана – це середнє значення в наборі даних після впорядкування списку значень (сортування). Медіана не чутлива до цих крайніх значень. Оскільки загальна кількість значень та фактичні значення у наборі даних однакові, середина у списку чи медіана залишається однаковою. Залежно від кількості випадаючих в наборі даних, середня величина або середня величина "перекошується" або змінюється в ту чи іншу сторону.

У той час як середнє значення час використовується для опису багатьох розподілів, воно залишає важливу частину загальної картини, яка є мінливістю розподілу. Наприклад, ми знаємо, що зовнішні значення можуть спотворювати середнє значення. Медіана наближає нас до того, що є головним у розподілі, проте ми все ще не знаємо, наскільки розподілені значення у вибірці. Основний спосіб опису змінності у вибірці – це обчислення різниці між найвищим і найнижчим значенням для змінної. Ця статистика відома як діапазон.

Дисперсія розподілу – це міра того, наскільки кожне значення в наборі даних від середнього. З дисперсією пов'язане стандартне відхилення, яке використовується для стандартизації розподілів як частини нормальної кривої, як показано на рис.5.3.

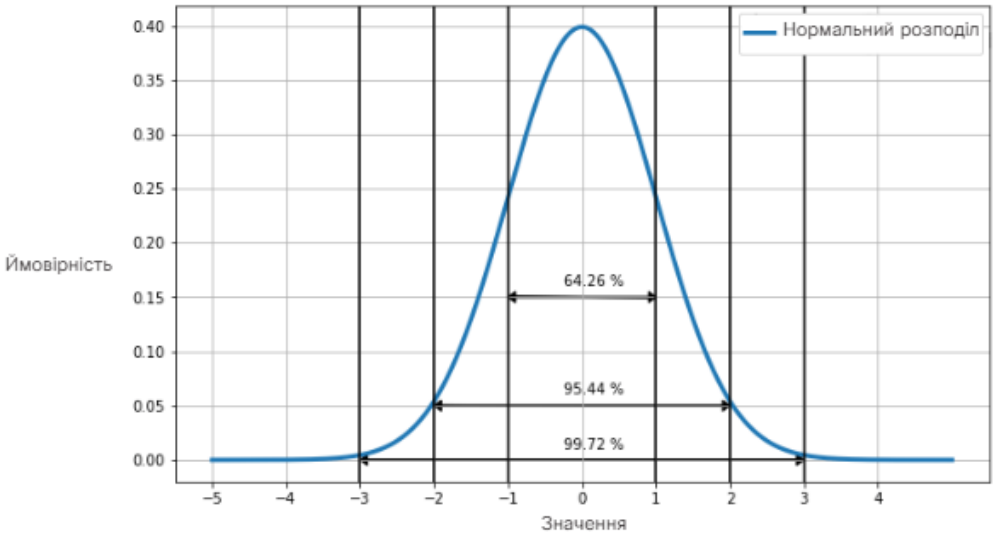


Рис. 5.3. Стандартні відхилення вибірки [1]

5.2. Використання Pandas

Pandas – це бібліотека з відкритим кодом для Python, яка додає високопродуктивні структури даних та інструменти для аналізу великих наборів даних. Структури даних Pandas включають структури рядів (series) та структури дата фреймів (dataframe). Дата фрейми є первинною структурою Pandas і є найбільш часто використовуваними. Дата фрейм схожий на електронну таблицю з рядками та стовпцями. Крім того, фрейми даних можуть мати додаткові індекси та стовпці, які є мітками для рядків та стовпців [2].

Дата фрейми легко будуються з ряду інших структур даних та зовнішніх файлів, таких як csv. Об'єктам кадрів даних доступний широкий спектр методів. Рядки та стовпці можуть маніпулювати різними способами, і оператори доступні для виконання математичних, рядкових та логічних перетворень на вміст фрейму даних (рис. 5.4).

	First Name	Last Name	Phone Number
1	Mary	Pratt	410-555-9697
2	Oscar	Milde	555-887-9547
3	Timmy	Thomas	471-555-9687

- two-dimensional
- labeled
- different types accomodated

Рис. 5.4. Компоненти дата фрейму Pandas [1]

Pandas імпортується в програму Python, використовуючи **import**, як і інші модулі. Зазвичай для використання імпорту **pandas as pd** для посилання на компоненти pandas простіше набрати. Нижче наведено код, необхідний для створення дата фрейму, який зображений на рис. 5.5.

```
import pandas as pd

data = {'First Name': ['Mary', 'Oscar', 'Timmy'],
        'Last Name': ['Pratt', 'Milde','Thomas'],
        'Phone Number': ['410-555-9697', '555-887-9547', '471-555-9687']}
directory = pd.DataFrame(data, columns=['First Name', 'Last Name', 'Phone Number'])
```

Directory

	First Name	Last Name	Phone Number
0	Mary	Pratt	410-555-9697
1	Oscar	Milde	555-887-9547
2	Timmy	Thomas	471-555-9687

Рис. 5.5. Створення дата фрейму вручну [1]

5.3. Імпорт даних з файлів

Великі набори даних збираються з різних джерел і можуть існувати у вигляді файлів різного виду. Створення дата фрейму Pandas шляхом кодування значень даних окремо не дуже корисно для аналізу великих даних.

Pandas включає деякі дуже прості у використанні функції для імпорту даних із зовнішніх файлів, таких як csv, у дата фрейми. Ми відтворимо дата фрейми телефонного каталогу, цього разу з більшого файлу csv. Pandas включає в себе функцію дата фрейму, який називається **read_csv ().**

Процедура така:

Крок 1. Імпортуйте модуль Pandas.

```
import pandas as pd
```

Крок 2. Перевірте, чи файл доступний у поточній робочій директорії. У цьому випадку команда **head** Linux використовується для перевірки файлу та попереднього перегляду його вмісту.

```
!head -n 5 directory.csv
```

```
First Name,Last Name,Phone Number
Mary,Pratt,410-555-9697
Oscar,Milde,555-887-9547
Timmy,Thomas,471-555-9687
John,Smith,252-959-8421
```

Крок 3. Щоб імпортувати файл в об'єкт дата фрейм, використовуйте метод Pandas **read_csv ().**

```
df_directory = pd.read_csv('directory.csv')
```

Крок 4. Використовуйте метод дата фрейму pandas **info ()**, щоб переглянути короткий зміст файлу.

```
df_directory.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15 entries, 0 to 14
Data columns (total 3 columns):
First Name      15 non-null object
Last Name       15 non-null object
Phone Number    15 non-null object
dtypes: object(3)
```

Крок 5. Відображення дата фрейму. Метод **head ()** використовується для відображення заголовків, індексу та значень для перших п'яти рядків.

```
df_directory.head()
```

Dataframe records

	First Name	Last Name	Phone Number
0	Mary	Pratt	410-555-9697
1	Oscar	Milde	555-887-9547
2	Timmy	Thomas	471-555-9687

5.4. Імпорт даних з мережі Інтернет

Імпортувати дані з Інтернету за допомогою Pandas дуже просто. Хоча для доступу до веб-даних доступно багато інтерфейсів прикладних програм (API), включаючи потокову передачу даних, статичні набори даних також можна отримати з Інтернету на основі URL-адреси файлу.

У прикладі, показаному на рис. 5.6, набір даних імпортується в набір даних із великої колекції Гуманітарної біржі даних [3]. Цей веб-сайт є гарним ресурсом для людей, зацікавлених у вивченні даних, пов'язаних з міжнародними гуманітарними проблемами.

```
import pandas as pd

url =
'http://manage.humdata.org/hdx/api/exporter/indicator/csv/TT014/source/mdgs/fromYear
/1950/toYear/0/language/en/TT014_Baseline.csv'

from_url = pd.read_table(url,sep=',')

from_url.head()
```

Country Data from CSV

	Country Code	Country Name	2015	2014	2013	2012	2011	2010	2009	2008	...	19
0	AFG	AFGHANISTAN	27.7	27.7	27.7	27.7	27.7	27.3	27.7	27.7	...	Na
1	AGO	ANGOLA	36.8	36.8	34.1	38.2	38.6	38.6	37.3	15.0	...	15
2	ALB	ALBANIA	20.7	20.0	15.7	15.7	16.4	16.4	7.1	7.1		Na

5 rows x 28 columns

```
from_url.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 192 entries, 0 to 191
Data columns (total 28 columns):
Country code    192 non-null object
Country name    192 non-null object
```

Рис. 5.6. Імпорт даних з мережі Інтернет у Pandas [1]

Ми імпортуємо набір даних, що містять інформацію про відсоток жінок, які працюють у національних парламентах для ряду країн протягом певних років. Процес включає в себе наступні кроки:

Крок 1. Імпорт Pandas.

Крок 2. Створення об'єкта рядка, який містить URL-адресу файлу.

Крок 3. Імпорт файлу в об'єкт дата фрейм методом pandas **read_table ().**

Крок 4. Перевірте імпорт за допомогою **head ()** та **info ()**. Висновок **info ()** вказує на кількість відсутніх значень (нульових записів), що є різницею між загальною кількістю записів та кількістю ненульових записів за кожен рік.

В мережі Інтернет є багато джерел даних. Наприклад, такі сайти, як Google і Twitter, мають API, які дозволяють підключати програми Python до потокових даних.

5.5. Описова статистика в Pandas

Pandas забезпечує дуже простий спосіб перегляду основних описових статистичних даних для фрейму даних. Метод **describe()** для об'єктів дата фрейму відображає наступне для числових типів даних:

- **count** – кількість значень, включених до статистики;
- **mean** – середнє значення;
- **std** – стандартне відхилення розподілу;
- **min** – найменше значення в розподілі;
- **25%** – значення для першого квартилю (25% значень знаходяться на рівні цього значення або нижче);
- **50%** – значення для другого квартилю або медіани (50% значень знаходяться на рівні цього значення або нижче);
- **75%** – значення для другого квартилю (75% значень знаходяться на рівні цього значення або нижче);
- **max** – найвище значення в розподілі.

5.6. Засоби для кореляційного аналізу в Pandas

Причинно-наслідкові зв’язки та кореляція – це типи зв’язків між умовами чи подіями. Причинно-наслідковий зв’язок – це відносини, у яких одна річ змінюється або створюється безпосередньо через щось інше.

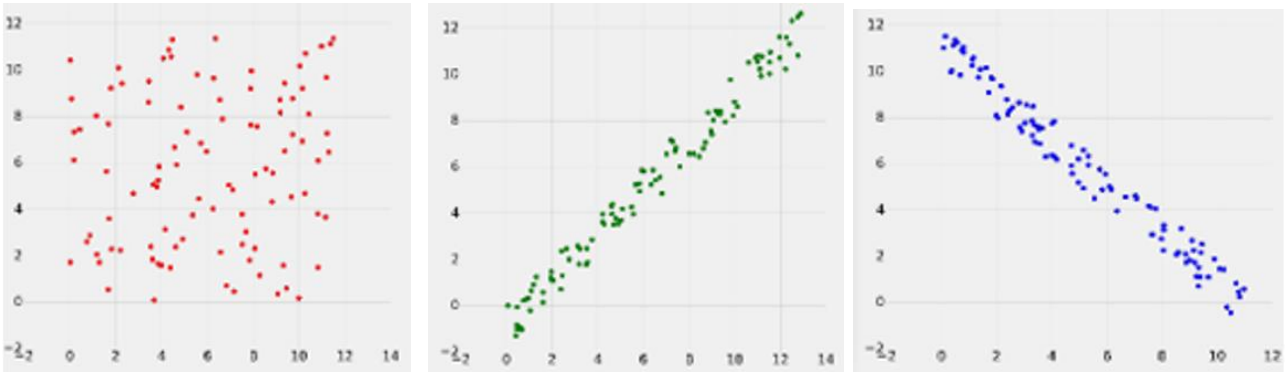
Наприклад, підвищення глобальної температури викликає зниження арктичної крижаної шапки. Це інтуїтивне відношення до явищ. Підвищення глобальної температури також може призвести до зменшення споживання вовни для використання у створенні теплого одягу. Чим тепліший клімат, тим менше буде попит на теплий одяг.

Кореляція – це зв’язок між величинами, у яких дві або більше величин змінюються з однаковою швидкістю. Наприклад, зменшуються глобальна температура та споживання вовни, ці величини змінюються з однаковою швидкістю та в аналогічному напрямку (обидві зменшуються).

Кореляції можуть бути позитивними та негативними. Позитивно корельовані величини змінюються в одному напрямку. Якщо одна величина збільшується, інша теж збільшується. Негативна кореляція виникає, коли величини змінюються в деякій схожій пропорції, але в протилежних напрямках. Іншими словами, якщо одна збільшується, інша зменшується аналогічно. Кореляції між величинами можна кількісно визначити, використовуючи статистичні підходи.

Найпоширенішою статистикою для обчислення кореляції є коефіцієнт кореляції Пірсона, це величина, яка виражається як значення між -1 і 1. Позитивні значення виражають позитивну залежність між змінами двох величин. Негативні значення виражають зворотну залежність.

Величина або позитивних, або негативних значень вказує на ступінь кореляції. Чим ближче значення до 1 або -1, тим міцніше зв'язок, 0 вказує на відсутність кореляції (рис. 5.7).



а) б) в)

Рис. 5.7 Розкиди невеликих наборів даних, які мають низьку, позитивну та негативну кореляцію, а) низька кореляція, r=0,0114, б) сильна позитивна кореляція, r=0,99, в) сильна негативна кореляція, r=-0,985 [1]

Метод **corr ()** простий у використанні. Розглянемо невеликий набір даних із більшого набору даних, який описує демографічні показники чисельності населення ряду міст.

```
import pandas as pd
csvName = 'povt-unemp.csv'
newFrame = pd.read_csv(csvName)
newFrame.head()
```

Poverty and Unemployment

	Poverty	%unempl
0	0.19	0.27
1	0.24	0.27

Дані спрощено, щоб містити лише два поля: відсоток людей, які живуть нижче межі бідності грошового доходу, і відсоток людей, які не працюють. Не дивно, що ці два поля повинні демонструвати сильну кореляцію. Це означало б, що ми очікували б, що для міста з великою кількістю людей, які живуть у злиднях, рівень безробіття також буде високим. У прикладі файл CSV, що містить дані, імпортується у дата фрейм. Дані швидко перевіряються за допомогою методів **head ()** та **describe()**, щоб переконатись, що імпорт працює належним чином. Для дата фрейму викликається метод **corr ()**. Результат відображається у таблиці кореляції. Бачимо, що з коефіцієнтом кореляції 0,73 безробіття має сильний зв’язок із бідністю.

```
newFrame.describe()
```

describe() Method Results

	Poverty	%unempl
count	81.000000	81.000000
mean	0.292099	0.362222

```
correl = newFrame.corr()

correl
```

corr() Method Results

	Poverty	%unempl
Poverty	1.000000	0.729233

Кореляції можна обчислити для кількох змінних одночасно. Це призведе до обчислення коефіцієнтів кореляції між усіма полями, що подаються в дата фрейм. Результатом може бути велика таблиця коефіцієнтів кореляції. Візуалізація, що називається **тепловою картою (heat map)**, корисна для розуміння того, як значення коефіцієнтів кореляції співвідносяться між собою. Графіки розсіювання корисні для швидкої візуалізації можливої кореляції у наборі даних. На тепловій карті поля в даних утворюють горизонтальну та вертикальну мітки для сітки значень (рис. 5.8).

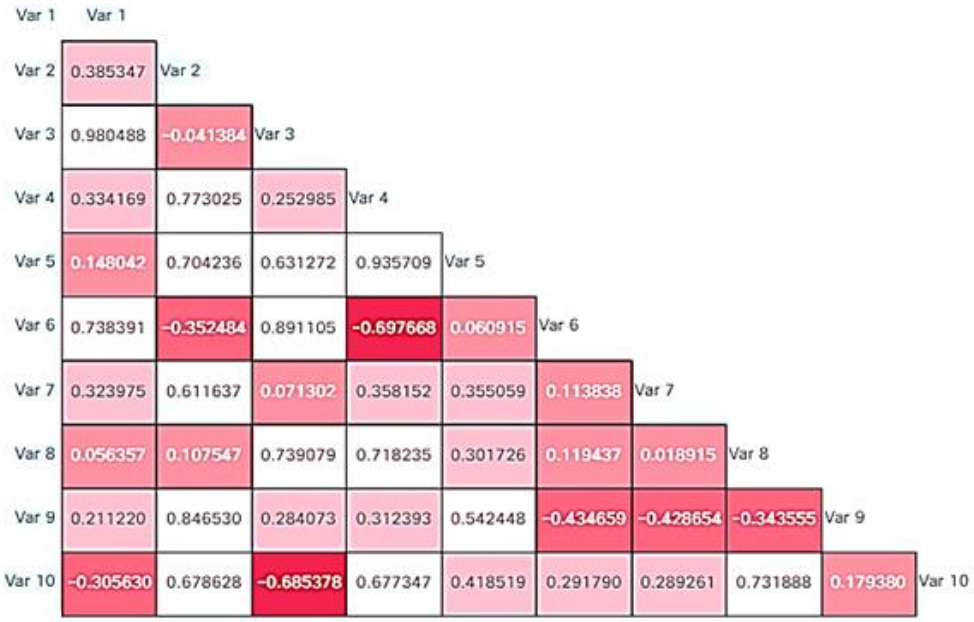


Рис. 5.8. Теплова карта коефіцієнтів кореляції між 10 змінними [1]

Кожне значення комірки є коефіцієнтом кореляції поля на горизонтальній розмірності сітки з полем на вертикальній осі.

Значення на перетині обраних розмірів є коефіцієнтом для цієї пари значень. Для інтерпретації даних значення кореляції мають кольорове кодування. Інтенсивність або відтінок кольору для кожного значення пропорційні цьому значенню. Наприклад, усі негативні коефіцієнти кореляції можуть бути представлені в червоному відтінку, а усі позитивні – у відтінку синього. Чим глибший колір, тим ближче значення до 1 або -1. Це допомагає зрозуміти значення з даних кореляції.

Висновок до лекції 5

Дослідницький аналіз даних дає описові та графічні узагальнення даних з для того, щоб з результатів виявити цікаві закономірності. Спостереження, змінні та значення мають вирішальне значення для аналізу.

Pandas – це бібліотека Python з відкритим кодом з інструментами для аналізу великих наборів даних, імпорту даних із файлів та з мережі Інтернет, перегляду описової статистики, пошуку статистичних залежностей для наборів даних. Дані зазвичай потребують очищення, перетворення та оброблення перед аналізом даних.

Питання для закріплення

1. Які статистичні підходи до аналітики великих даних ви знаєте?
2. Яке призначення бібліотеки Pandas?
3. Як відбувається імпорт даних з файлів у Pandas?
4. Як відбувається імпорт даних з мережі Інтернет?
5. Як перевірити описову статистику даних у Pandas?
6. Які засоби для кореляційного аналізу в Pandas ви знаєте?

Список рекомендованої літератури

1. IoT Fundamentals: Big Data & Analytics // Електронний ресурс. Режим доступу: https://www.netacad.com/courses/iot/big-data-analytics

2. Pandas // Електронний ресурс. Режим доступу: https://github.com/pandas-dev/pandas

3. The Humanitarian Data Exchange // Електронний ресурс. Режим доступу: https://data.humdata.org/

4. Pandas // Електронний ресурс. Режим доступу: https://www.w3schools.com /python/pandas/pandas_intro.asp