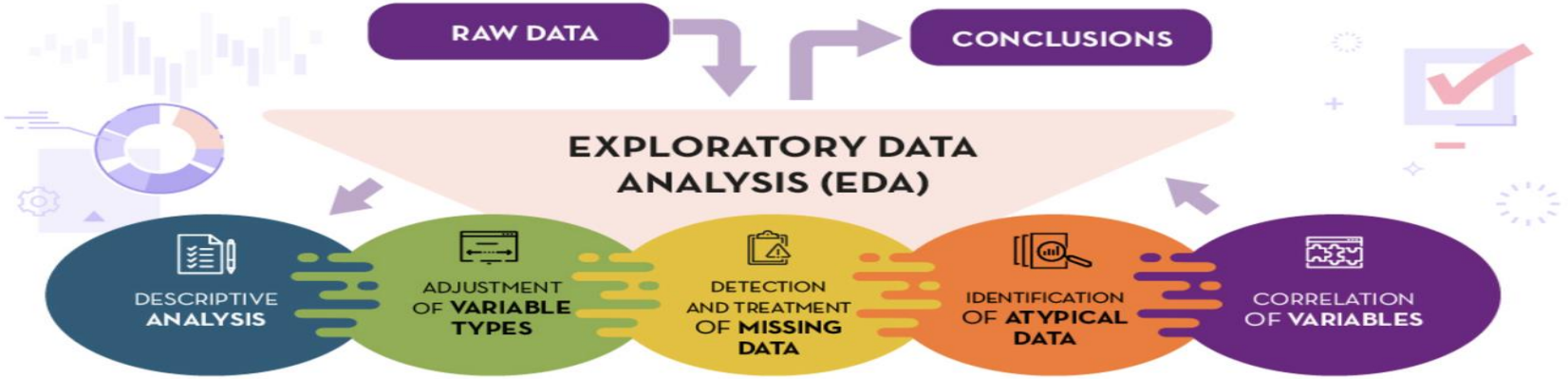
Методи розвідувального аналізу (EDA) та підготовки даних



Задачі розвідувального аналіза даних (EDA)

Крок 1. Загальний огляд даних - розуміння загальної структури та якості даних

```
yield_df = pd.read_csv('yield_df.csv', index_col=0)
```

+ Code + Text

```
yield_df.head()
```

	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_
0	Albania	Maize	1990	36613	14
1	Albania	Potatoes	1990	66667	14

```
yield_df.describe()
```

	Year	hg/ha_yield	average_rain_fall_mm_per_year		
count	28242.000000	28242.000000	28242.00000		
mean	2001.544296	77053.332094	1149.05598		
std	7.051905	84956.612897	709.81215		
min	1990.000000	50.000000	51.00000	0.040000	1.300000
25%	1995.000000	19919.250000	593.00000	1702.000000	16.702500
50%	2001.000000	38295.000000	1083.00000	17529.440000	21.510000

```
[17] yield_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 28242 entries, 0 to 28241
Data columns (total 7 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Area                           28242 non-null  object
 1   Item                           28242 non-null  object
 2   Year                           28242 non-null  int64
 3   hg/ha_yield                    28242 non-null  int64
 4   average_rain_fall_mm_per_year  28242 non-null  float64
 5   pesticides_tonnes              28242 non-null  float64
 6   avg_temp                       28242 non-null  float64
dtypes: float64(3), int64(2), object(2)
memory usage: 1.7+ MB
```

Загальний огляд даних: ProfileReport

```
from ydata_profiling import ProfileReport
profile = ProfileReport(yield_df, title="Pandas Profiling Report")
profile
```

Summarize dataset: 100% 41/41 [00:10<00:00, 2.66it/s, Completed]

Generate report structure: 100% 1/1 [00:06<00:00, 6.03s/it]

Render HTML: 100% 1/1 [00:01<00:00, 1.31s/it]

Pandas Profiling Report

Overview Variables Interactions Correlations Missing values Sample Duplicate row

Загальний огляд даних: ProfileReport

Overview

Overview | Alerts 1 | Reproduction

Dataset statistics

Number of variables	7
Number of observations	28242
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	2101
Duplicate rows (%)	7.4%
Total size in memory	1.7 MiB
Average record size in memory	64.0 B

Variable types

Text	1
Categorical	1
Numeric	5

Overview | Alerts 1 | Reproduction

Alerts

Dataset has 2101 (7.4%) duplicate rows

Загальний огляд даних: ProfileReport - аналіз змінних

Variables

Area

Area

Text

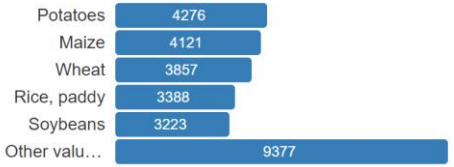
Distinct	101
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	441.3 KiB

Item

Item

Categorical

Distinct	10
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	441.3 KiB



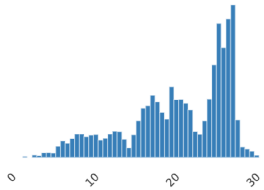
Загальний огляд даних: ProfileReport - аналіз змінних

avg_temp

avg_temp

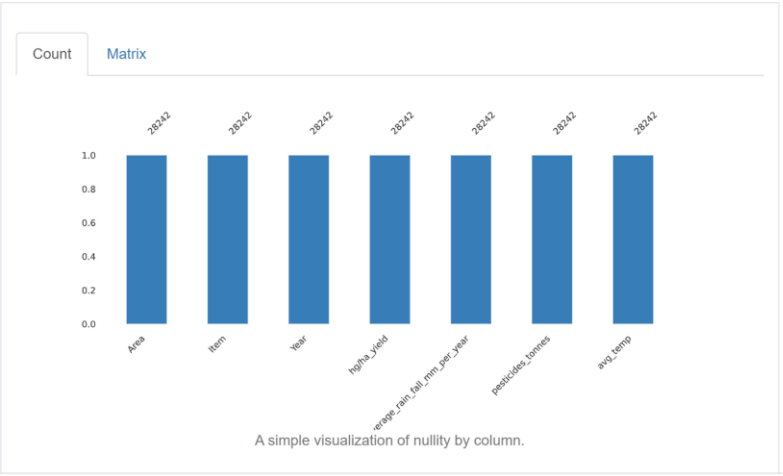
Real number (ℝ)

Distinct	1831	**Minimum**	1.3
Distinct (%)	6.5%	**Maximum**	30.65
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	20.542627	**Memory size**	441.3 KiB



More details

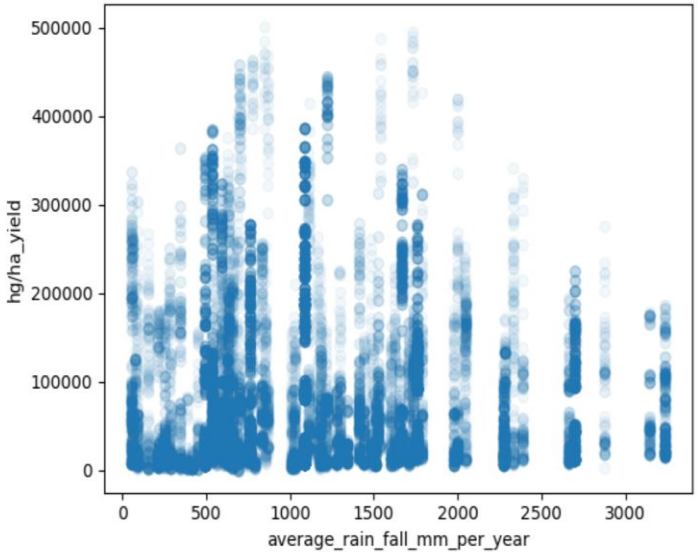
Missing values

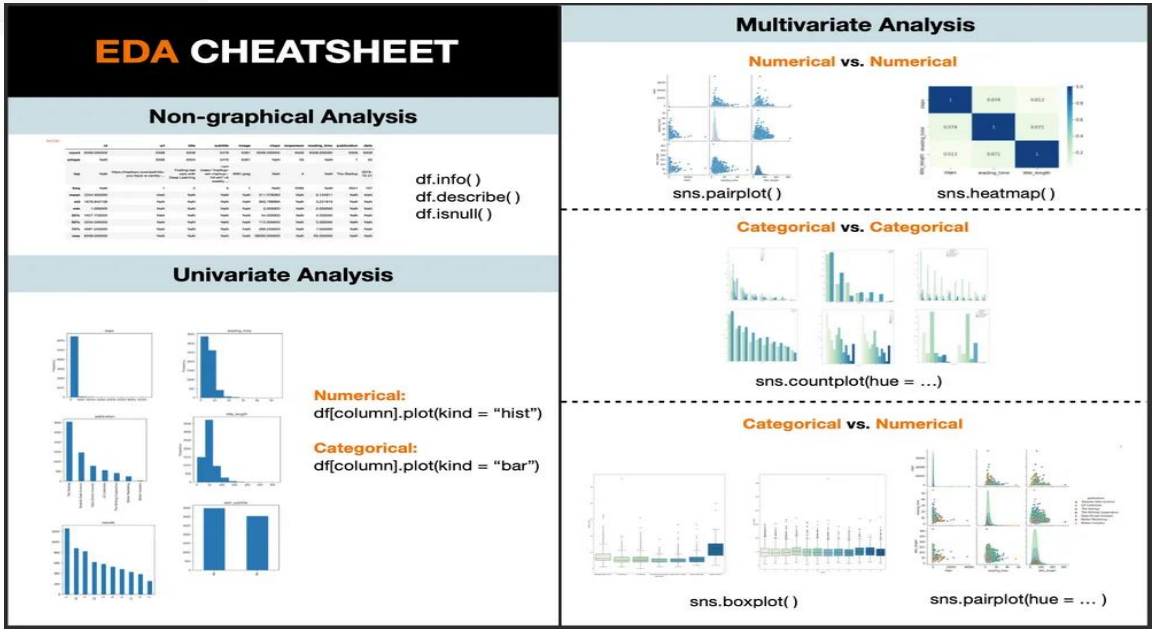


Загальний огляд даних: візуалізація залежностей

```
import matplotlib.pyplot as plt

plt.scatter(x=yield_df_all['average_rain_fall_mm_per_year'], y=yield_df_all['hg/ha_yield'], alpha=0.05)
plt.xlabel('average_rain_fall_mm_per_year')
plt.ylabel('hg/ha_yield')
plt.show()
```





Крок 2. Аналіз та виправлення пропущених даних, способи усунення

sklearn.impute.SimpleImputer

class sklearn.impute.**SimpleImputer**(*, *missing_values=nan*, *strategy='mean'*, *fill_value=None*, *copy=True*, *add_indicator=False*, *keep_empty_features=False*)

missing_values ***int, float, str, np.nan, None or pandas.NA, default=np.nan***

strategy, *default='mean'* - replace missing values using the mean along each column.
"median" - replace missing values using the median along each column.
"most_frequent" - replace missing using the most frequent value along each column.
"constant" - replace missing values with fill_value.

sklearn.impute.KNNImputer

class sklearn.impute.**KNNImputer**(*, *missing_values=nan*, *n_neighbors=5*, *weights='uniform'*, *metric='nan_euclidean'*, *copy=True*, *add_indicator=False*, *keep_empty_features=False*)

missing_values*int, float, str, np.nan or None, default=np.nan*
n_neighbors*int, default=5* Number of neighboring samples to use for imputation.

weights*{'uniform', 'distance'} or callable, default='uniform'* Weight function used in prediction.
metric*{'nan_euclidean'} or callable, default='nan_euclidean* Distance metric for searching neighbors.

Крок 3. Виявлення та обробка викидів та аномалій

Методи виявлення:

1. Правило **3 Сигма** - якщо дані розподілені нормально
2. Правило **Inter-Quartile Range (IQR)**:
   (Q1 – 1.5 IQR, Q3 + 1.5 IQR), де IQR = Q3 - Q1
3. a-Percentile:
   (1% percentile, 99% percentile)

Методи обробки:

Trimming - ***видалення*** - спостереження із викидами видаляються із аналізу.
Capping - ***обмеження*** - значення, що виходять за межі певних граничних значень замінюються на ці граничні або якісь наперед задані значення.

Discretization - ***дискретизація*** - заміна фактичних даних на групи.

Крок 4. **Пошук зв'язків та кореляцій** - встановлення та візуалізація взаємозв'язків між різними змінними та факторами.

Correlations

Auto

Heatmap | Table

	Item	Year	average_rain_fall_mm_per_year	avg_temp	hg/ha_yield	pesticides_tonnes
Item	1.000	0.001	-0.059	-0.082	-0.242	0.064
Year	0.001	1.000	-0.004	0.026	0.093	0.028
average_rain_fall_mm_per_year	-0.059	-0.004	1.000	0.291	0.075	0.139
avg_temp	-0.082	0.026	0.291	1.000	-0.105	-0.126
hg/ha_yield	-0.242	0.093	0.075	-0.105	1.000	0.156
pesticides_tonnes	0.064	0.028	0.139	-0.126	0.156	1.000

Питання?