Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет інформатики та обчислювальної техніки
Кафедра інформаційних систем та технологій

Лабораторна робота №6

з дисципліни

Аналіз даних з використанням мови Python

Виконав:

студент групи ІА-24:
Криворучек В.С.

Перевірила:

ст. викладач
Тимофєєва Ю.С.

Київ 2025

Тема: Попередня обробка даних в Pandas

Мета роботи: Ознайомитись з операціями попередньої обробки даних Pandas.

Хід роботи

Завдання:

Файл Version 10.xlsx

Створити програму, яка виконує наступні завдання, використовуючи файл відповідно до варіанту:

1. Читає файл та змінює назви стовпців.
2. Знаходить проблеми з даними та виконує попередню обробку даних для усунення цих проблем.

Код програми:

```
import pandas as pd
from fuzzywuzzy import process
import numpy as np

# === 1. Завантаження Excel-файлу ===
df = pd.read_excel('Version 10.xlsx')

# === 2. Перейменування стовпців ===
df.rename(columns={
    'ID': 'ID',
    'Warehouse_block': 'Warehouse',
    'Mode_of_Shipment': 'Shipment_Mode',
    'Customer_care_calls': 'Care_Calls',
    'Customer_rating': 'Rating',
    'Cost_of_the_Product': 'Cost',
    'Prior_purchases': 'Purchases',
    'Product_importance': 'Importance',
    'Gender': 'Gender',
    'Discount_offered': 'Discount',
    'Weight_in_gms': 'Weight',
    'Reached.on.Time_Y.N': 'DeliveredOnTime'
}, inplace=True)

# === 3. Функція для виправлення категоріальних значень ===
def correct_spelling_fuzzy(series, valid_values, min_score=70):
    corrected = []
    for value in series:
        str_value = str(value).strip()
        if not str_value or str_value in ['???', '?', 'nan', 'NaN',
'None']:
            corrected.append(np.nan)
        else:
            match, score = process.extractOne(str_value, valid_values)
            corrected.append(match if score >= min_score else value)
```

```
    return pd.Series(corrected)

# === 4. Виправлення орфографічних помилок у категоріях ===
valid_mode = ['Flight', 'Ship', 'Road']
valid_importance = ['low', 'medium', 'high']
valid_gender = ['M', 'F']
valid_warehouse = ['A', 'B', 'C', 'D', 'F']

df['Shipment_Mode'] = correct_spelling_fuzzy(df['Shipment_Mode'],
valid_mode)
df['Importance'] = correct_spelling_fuzzy(df['Importance'],
valid_importance)
df['Gender'] = correct_spelling_fuzzy(df['Gender'], valid_gender)
df['Warehouse'] = correct_spelling_fuzzy(df['Warehouse'],
valid_warehouse)

# === 5. Попередня обробка числових і текстових значень ===

# Видаляємо дублікати
df.drop_duplicates(inplace=True)

# Заповнюємо числові NaN середніми значеннями
df.fillna({col: df[col].mean() for col in
df.select_dtypes(include='number')}, inplace=True)

# Заповнюємо текстові NaN найчастішим значенням
df.fillna({col: df[col].mode()[0] for col in
df.select_dtypes(include='object')}, inplace=True)

# === 6. Збереження результату ===
df.to_excel('cleaned_output.xlsx', index=False)

# === 7. Вивід результатів ===
print("\nПерші 5 рядків:")
print(df.head().to_string(index=False))
```

У результаті отримуємо виправлений файл:

Unnamed: 0	ID	Warehouse	Shipment_Mode	Care_Calls	Rating	Cost	Purchases	Importance	Gender	Discount	Weight	Deliver
0	1	D	Flight	4	2	177	3	low	F	44	1233	1
1	2	F	Flight	4	5	216	2	low	M	59	3088	1
2	3	A	Flight	2	2	183	4	low	M	48	3374	1
3	4	B	Flight	3	3	176	4	medium	M	10	1177	1
4	5	C	Flight	2	2	184	3	medium	F	46	2484	1
5	6	F	Flight	3	1	162	3	medium	F	12	1417	1
6	7	D	Flight	3	4	250	3	low	F	3	2371	1
7	8	F	Flight	4	1	233	2	low	F	48	2804	1
8	9	A	Flight	3	4	150	3	low	F	11	1861	1
9	10	B	Flight	3	2	164	3	medium	F	29	1187	1
10	11	C	Flight	3	4	189	2	medium	M	12	2888	1
11	12	F	Flight	4	5	232	3	medium	F	32	3253	1
12	13	D	Flight	3	5	198	3	medium	F	1	3667	1
13	14	F	Flight	4	4	275	3	high	M	29	2602	1
14	15	A	Flight	4	3	152	3	low	M	43	1009	1
15	16	B	Flight	4	3	227	3	low	F	45	2707	1
16	17	C	Flight	3	4	143	2	medium	F	6	1194	1
17	18	F	Ship	5	5	227	3	medium	M	36	3952	1
18	19	D	Ship	5	5	239	3	high	M	18	2495	1
19	20	F	Ship	4	5	145	3	medium	M	45	1059	1
20	21	A	Ship	3	3	161	2	medium	F	38	1521	1
21	22	B	Ship	3	1	232	4	medium	F	51	2899	1
22	23	C	Ship	2	5	156	2	low	M	2	1750	1
23	24	F	Ship	4	3	211	3	high	M	12	3922	1
24	25	D	Ship	4	5	251	2	medium	F	28	3561	1
25	26	F	Ship	3	1	225	4	low	M	29	3496	1

Висновок: У ході виконання даної лабораторної роботи я ознайомився з операціями попередньої обробки даних Pandas.