Міністерство освіти і науки України
Львівський національний університет імені Івана Франка
Факультет електроніки та комп'ютерних технологій
Кафедра системного проектування

Звіт
Про виконання лабораторної роботи №1
З курсу «Основи аналізу даних»
Вступ в Python для аналізу даних.

Виконав:
Студент групи ФЕІ-42
Прізвище Ім'я

Перевірив:
Асист. Азаров І.В.

Львів 2024

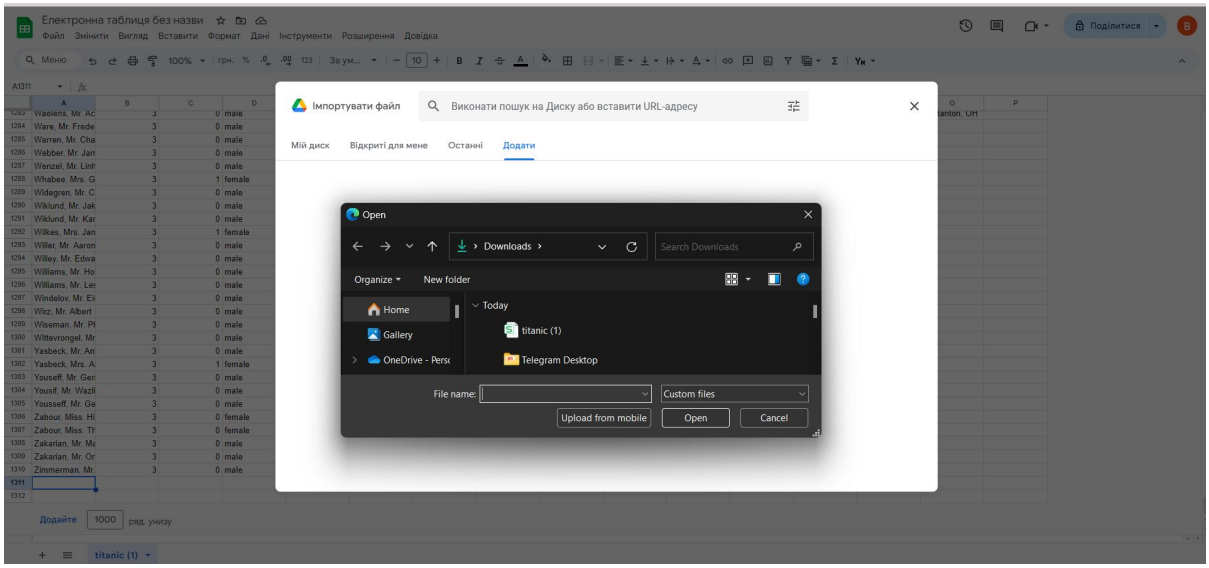
Мета: Ознайомитися з бібліотеками Python, що використовуються для аналізу даних.

Хід роботи:

Завдання

1. Робота з таблицями

- Імпортувати наявні дані



- Пояснити до якої шкали належать різні змінні в таблиці

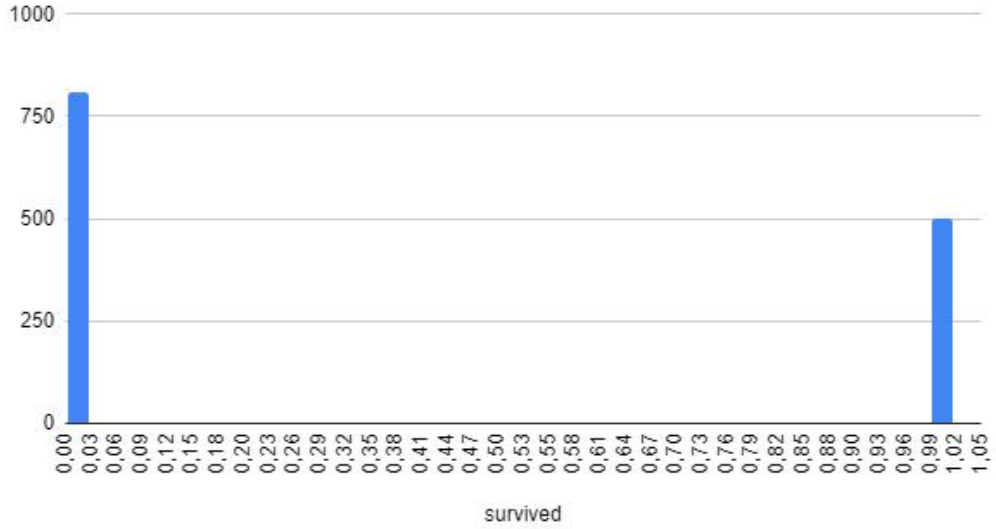
Змінні name,ticket,home.dest належать до якісних решта до кількісних

- Додайте вигадану особу з своїм прізвищем та збережіть зміни

1311	Reinhardt Vlady:	1	0	male	51	0	0	315757	7.2400		S	1	14 New York, NY

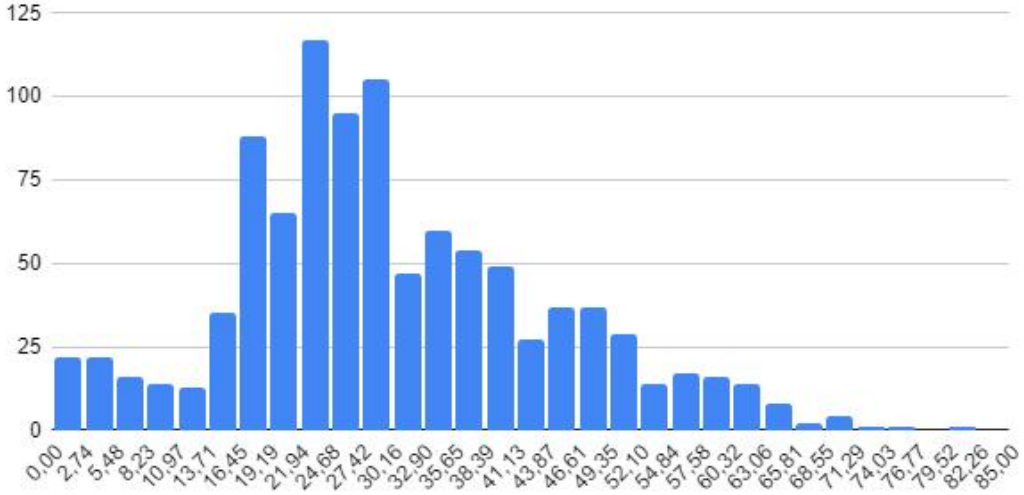
- Побудувати гістограми для кожного стовпця

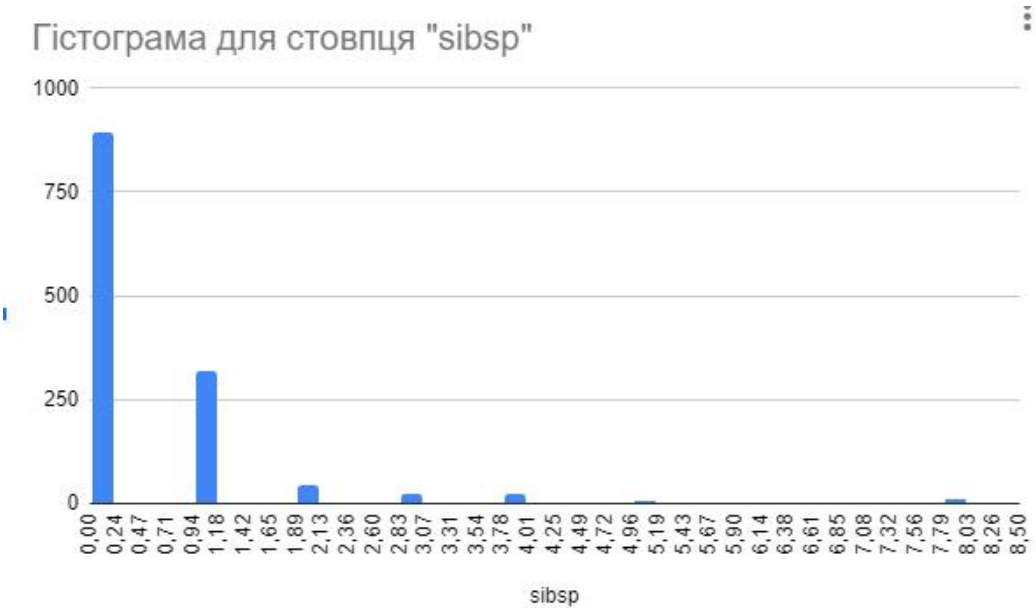
Гістограма для стовпця "survived"



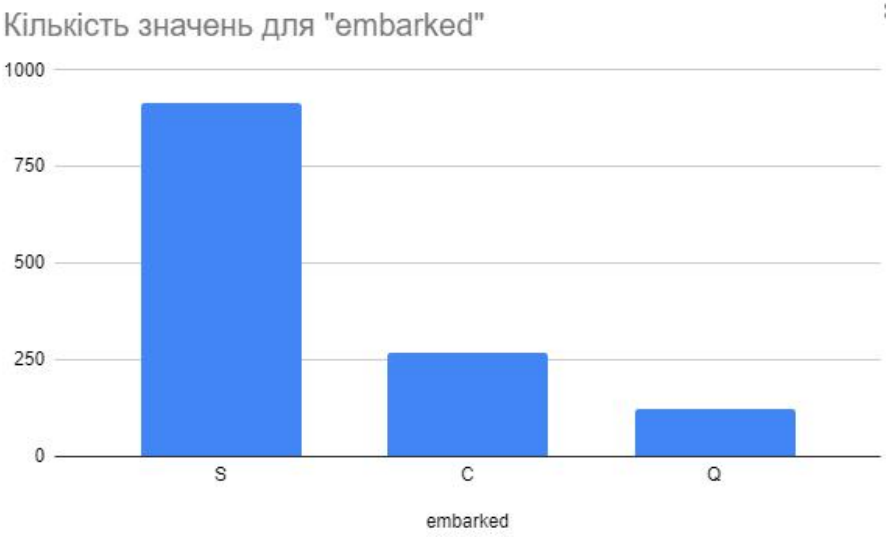
Гістограма для стовпця "Кількість значень для "gender""

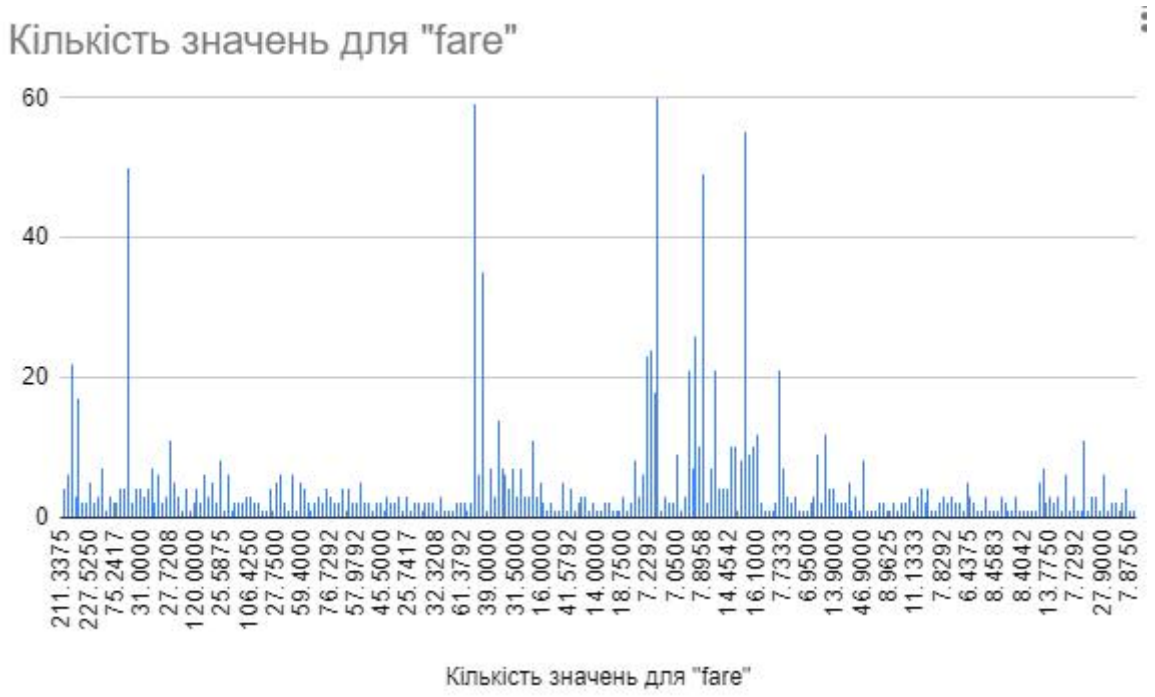
Гістограма для стовпця "age"




Гістограма для стовпця "sibsp"
1000
750
500
250
0
0,00 0,24 0,47 0,71 0,94 1,18 1,42 1,65 1,89 2,13 2,36 2,60 2,83 3,07 3,31 3,54 3,78 4,01 4,25 4,49 4,72 4,96 5,19 5,43 5,67 5,90 6,14 6,38 6,61 6,85 7,08 7,32 7,56 7,79 8,03 8,26 8,50
sibsp

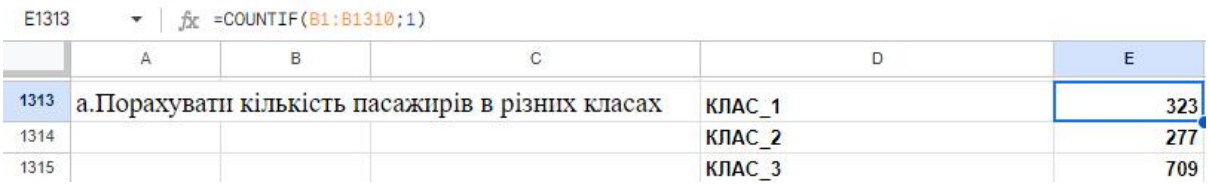

Гістограма для стовпця "parch"
1250
1000
750
500
250
0
0,00 0,26 0,53 0,79 1,06 1,32 1,58 1,85 2,11 2,38 2,64 2,90 3,17 3,43 3,69 3,96 4,22 4,49 4,75 5,01 5,28 5,54 5,81 6,07 6,33 6,60 6,86 7,13 7,39 7,65 7,92 8,18 8,44 8,71 8,97 9,24 9,50
parch


Кількість значень для "embarked"
1000
750
500
250
0
S
C
Q
embarked

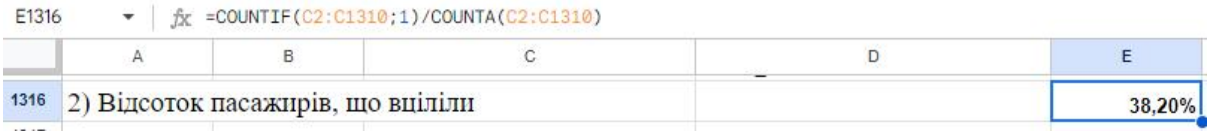


- Фільтрація даних за певними критеріями

a. Порахувати кількість пасажирів в різних класах



b. Відсоток пасажирів, що вціліли



c. Середній вік пасажирів

2. Підготовка середовища

- Встановити IDE Python

```
PS C:\Users\reung> python --version
Python 3.11.9
PS C:\Users\reung>
```

- Встановити NumPy, pandas, matlibplot

Всі бібліотеки були встановлені раніше.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

3. Виконання базових операцій з даними за допомогою Python (завантаження, фільтрація, нормалізація):

- Завантажити дані з CSV-файлу.

```
titanic_data = pd.read_csv("titanic.csv")
```

- Продемонструвати перших 7 пасажирів, перша літера прізвища яких співпадає з прізвищем автора

```
# Фільтрація пасажирів за прізвищем, яке починається з певної літери (наприклад, 'R')
titanic_data['surname'] = titanic_data['name'].str.split(',').str[0]

# Фільтруємо пасажирів за літерою 'R' і виводимо перші 7 записів
filtered_passengers = titanic_data[titanic_data['surname'].str.startswith('R')].head(7)
print(filtered_passengers)
```

[18] ✓ 0.0s

```
                                                  name  pclass  survived  \
234                    Reuchlin, Jonkheer. John George       1         0
235                 Rheims, Mr. George Alexander Lucien       1         1
236                               Ringhini, Mr. Sante       1         0
237                               Robbins, Mr. Victor       1         0
238  Robert, Mrs. Edward Scott (Elisabeth Walton Mc...       1         1
239                Roebling, Mr. Washington Augustus II       1         0
240        Romaine, Mr. Charles Hallace ("Mr C Rolmane")       1         1

     gender   age  sibsp  parch    ticket      fare cabin embarked boat  \
234    male  38.0      0      0     19972    0.0000   NaN        S  NaN
235    male   NaN      0      0  PC 17607   39.6000   NaN        S    A
236    male  22.0      0      0  PC 17760  135.6333   NaN        C  NaN
237    male   NaN      0      0  PC 17757  227.5250   NaN        C  NaN
238  female  43.0      0      1     24160  211.3375    B3        S    2
239    male  31.0      0      0  PC 17590   50.4958   A24        S  NaN
240    male  45.0      0      0    111428   26.5500   NaN        S    9
```

- Фільтрація даних за певними критеріями

a. Порахувати кількість пасажирів в різних класах

```
# Кількість пасажирів у кожному класі
passenger_count_by_class = titanic_data['pclass'].value_counts()
print("Кількість пасажирів у кожному класі:\n", passenger_count_by_class)
```

✓ 0.0s

```
Кількість пасажирів у кожному класі:
 pclass
3    709
1    324
2    277
```

b. Відсоток пасажирів, що вціліли

```
# Відсоток пасажирів, що вижили
survived_percentage = titanic_data['survived'].mean() * 100
print(f"Відсоток пасажирів, що вижили: {survived_percentage:.2f}%")
```

[10] ✓ 0.0s

```
Відсоток пасажирів, що вижили: 38.24%
```

c. Середній вік пасажирів

```
# Середній вік пасажирів (ігноруємо відсутні значення)
average_age = titanic_data['age'].mean()
print(f"Середній вік пасажирів: {average_age:.2f} років")
```

[10]

```
Середній вік пасажирів: 29.88 років
```

- Нормалізація даних.

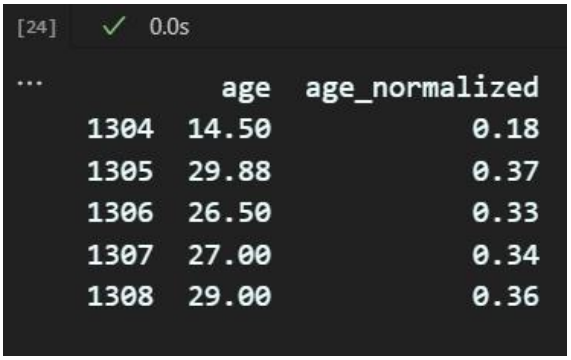
a. Виконати нормалізацію

```
# Обробка пропущених значень у стовпці 'age'
titanic_data['age'].fillna(titanic_data['age'].mean(), inplace=True)

# Нормалізація стовпця 'age'
titanic_data['age_normalized'] = (titanic_data['age'] - titanic_data['age'].min()) / (titanic_data['age'].max() - titanic_data['age']

# Заокруглення до двох знаків після коми
titanic_data['age_normalized'] = titanic_data['age_normalized'].round(2)
titanic_data['age'] = titanic_data['age'].round(2)

# Виведення останніх 5 рядків
print(titanic_data[['age', 'age_normalized']].tail())
```



b. Пояснити зміни в даних

Як правило, ми нормалізуємо дані, коли робимо якийсь тип аналізу, у якому ми маємо кілька змінних, виміряних у різних масштабах, і ми хочемо, щоб кожна зі змінних мала однаковий діапазон.

Це запобігає надмірному впливу однієї змінної, особливо якщо вона вимірюється в різних одиницях (тобто якщо одна змінна вимірюється в дюймах, а інша – у ярдах).

З іншого боку, ми зазвичай нормалізуємо дані, коли хочемо знати, скільки стандартних відхилень має кожне значення в наборі даних від середнього.

Нормалізований набір даних завжди матиме значення від 0 до 1. У нашому випадку, я нормалізував значення у колонці age, адже як можна побачити решта числових значень у таблиці в межах від 0 до 1.

Висновок: У процесі лабораторної роботи з аналізу даних було:

1. **Знайомство з Google Sheets:** було досліджено використання формул для фільтрації даних та побудова графіків (гістограм).
2. **Ознайомлено з бібліотеками Python** для роботи з даними (pandas, matplotlib).
3. **Використання Python:** імпортовано CSV-файл, додано вигадану особу, здійснено фільтрацію за класом пасажира, відсотком виживання та середнім віком

Отримані результати допомогли краще зрозуміти дані та підготувати їх для подальшого аналізу.