МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Інститут комп'ютерних наук та інформаційних технологій
Кафедра систем штучного інтелекту

Лабораторна робота №1

З курсу «Візуалізація даних»

Аналіз даних та статистичне виведення

Виконав:
ст. гр. КН-310
Бікєєв Андрій

Викладач:
Бойко Наталія Іванівна

Львів – 2020

1. Умова завдання

1. Завантажити дані та дослідити їх.
2. Переглянемо перші шість, перші п'ятнадцать та останні шість рядків з наших даних.
3. Дізнатися, яка кількість квартир продається у кожному місті згідно до нашого датасету.
4. Побудуємо стовпчикові діаграми для:
   (а) кількості кімнат
   (б) змінної площа
   (в) розподіл квартир, які продаються за загальною площею
5. Побудувати графік розсіювання, а саме залежності ціни від загальної площі.
6. Побудувати графік розподілу цін по містах.

2. Хід роботи

Для лабораторної роботи я використав бібліотеки:

- pandas - для імпорту .csv файлу у датафрейм.
- matpotlib - для будування графіків.
- seaborne - для будування більш "складних"графіків.

2.1. Імпортування бібліотек і даних

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

data = pd.read_csv("flats.csv", sep = ',', decimal=',')
data.columns = [col.replace('"', '') for col in data.columns]
```

2.2. Знаходження кількості вимірів, відображення і огляд інформації

Знайдіть кількість вимірів датафрейму flats.
Відобразіть перші шість рядків, перші п'ятнадцять рядків, останні шість рядків.
Відобразіть імена датафрейму.

Рис. 1: Завдання

```
print(data.shape)

print(data.head(6))
print(data.head(15))
print(data.tail(6))

print(data.columns)
```

```
(839, 4)
     Місто  Кімнат  Загальна_площа       Ціна
0  Вінниця       3           120.0  1875000.0
1  Вінниця       3            66.0   975000.0
2  Вінниця       2            66.0  1375000.0
3  Вінниця       2            44.0   637500.0
4  Вінниця       3            63.0   835000.0
5  Вінниця       1            31.0   562500.0
      Місто  Кімнат  Загальна_площа       Ціна
0   Вінниця       3           120.0  1875000.0
1   Вінниця       3            66.0   975000.0
2   Вінниця       2            66.0  1375000.0
3   Вінниця       2            44.0   637500.0
4   Вінниця       3            63.0   835000.0
5   Вінниця       1            31.0   562500.0
6   Вінниця       3            46.0  1150000.0
7   Вінниця       3            64.0   800000.0
8   Вінниця       1            35.0   424975.0
9   Вінниця       6           200.0    12500.0
10  Вінниця       2            46.0   500000.0
11  Вінниця       1            50.0   999975.0
12  Вінниця       1            38.0   512500.0
13  Вінниця       2            68.0  1000000.0
14  Вінниця       3            98.0  2575000.0
            Місто  Кімнат  Загальна_площа      Ціна
833  Хмельницький       1           35.58  212500.0
834  Хмельницький       1           52.00  330000.0
835  Хмельницький       1           41.00  325000.0
836  Хмельницький       1           47.00  375000.0
837  Хмельницький       2           53.00  387500.0
838  Хмельницький       2           60.00  522500.0
Index(['Місто', 'Кімнат', 'Загальна_площа', 'Ціна'], dtype='object')
```

Рис. 2: Результат виконання

2.3. Перевірка та попередній аналіз даних

Скільки змінних у наборі даних flats?
Яка кількість міст у наборі даних flats?
Чи всі з них дійсно є містами?
Яка кількість трикімнатних квартир продається в місті Одеса?
Яка медіана площі однокімнатної квартири в місті Львів?

Рис. 3: Завдання

```
print(len(data.columns))
print(len(data.filter(items="Місто")))
print(len(data[(data.Місто == "Одеса") & (data.Кімнат == 3)]))
```

```
data["Загальна_площа"] = data["Загальна_площа"].str.replace(
    ',', '.'
).astype(float)
newdata = data[(data.Місто == "Львів") & (data.Кімнат == 1)]
print(newdata["Загальна_площа"].median())
```

```
4
839
{'Запоріжжя', 'Миколаїв', 'Вінниця', 'Києво-Святошинський', 'Київ', 'Хмельницький', 'Ів
11
43.0
```

Рис. 4: Результат виконання

Як бачимо, ні, не всі записи у наборі даних flats є містами, наприклад Києво-Святошинський є районом Київської області, а не містом. Це може привести до спотворення даних, адже ці записи складають приблизно 2.2% від усіх записів(19/839).

2.4. Побудова діаграм, та їх аналіз

1. **Стовпчикова діаграма за кількістю кімнат.**

```
print(len(data.columns))
print(len(data.filter(items="Місто")))
print(len(data[(data.Місто == "Одеса") & (data.Кімнат == 3)]))

data["Загальна_площа"] = data["Загальна_площа"].str.replace(
    ',', '.'
).astype(float)
newdata = data[(data.Місто == "Львів") & (data.Кімнат == 1)]
print(newdata["Загальна_площа"].median())
```

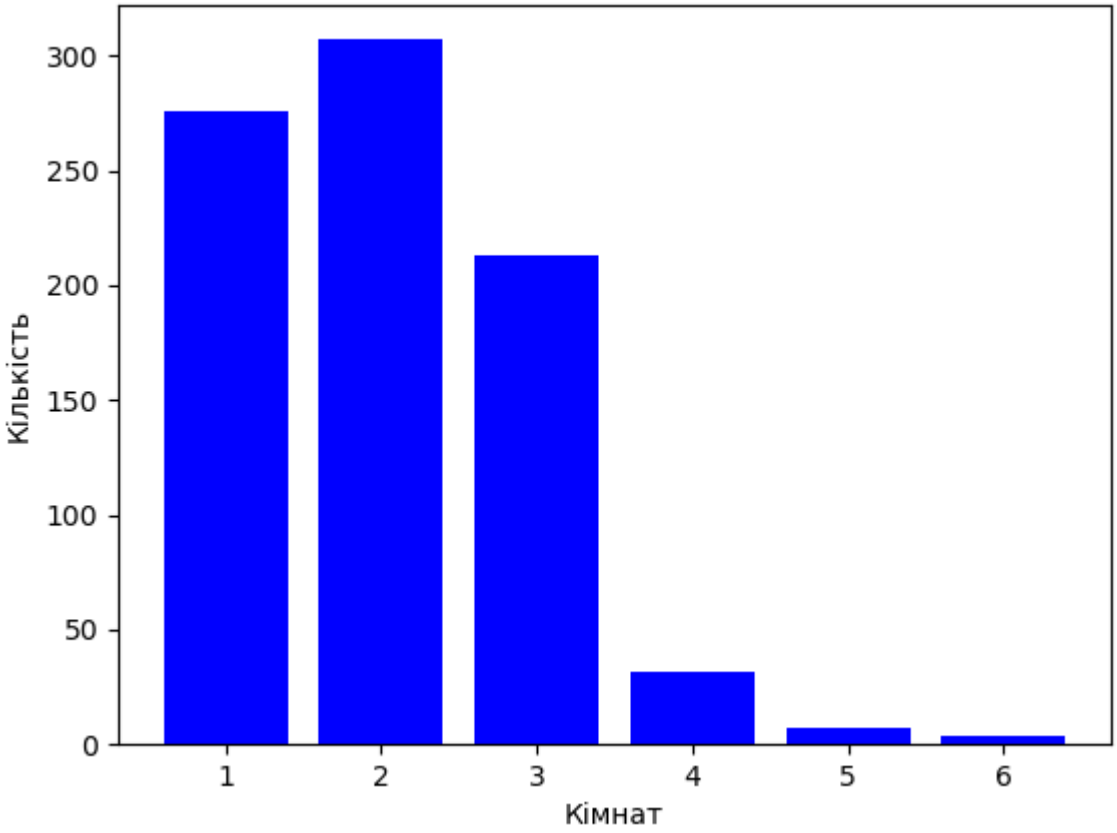

Рис. 5: Діаграма Стовпчикова "Кількість кімнат"

За допомогою діаграми, бачимо, що кількість кімнат у квартирах, які продають, розподілена нерівномірно, наприклад найчастіше зустрічаються квартири двухкімнатні(300 записів), що є модою нашої вибірки. Рідше всього продають квартири 6-и кімнатні, що певно зумовлено низьким попитом на такого роду квартири.

2. **Стовпчикова діаграма за загальною площею.**

```
values = data.groupby('Загальна_площа').size().reset_index(name='Кількість')
x_ax = values['Загальна_площа'].tolist()
y_ax = values['Кількість'].tolist()
plt.bar(x_ax,y_ax, color='r')
plt.xlabel('Загальна_площа')
plt.ylabel('Кількість')
plt.show()
```

На діаграмі за загальною площею можна побачити схожий розподіл на розподіл діаграми за кількістю кімнат, адже їх кількість корелює з площею, і деякі чинники в цих випадках спільні.

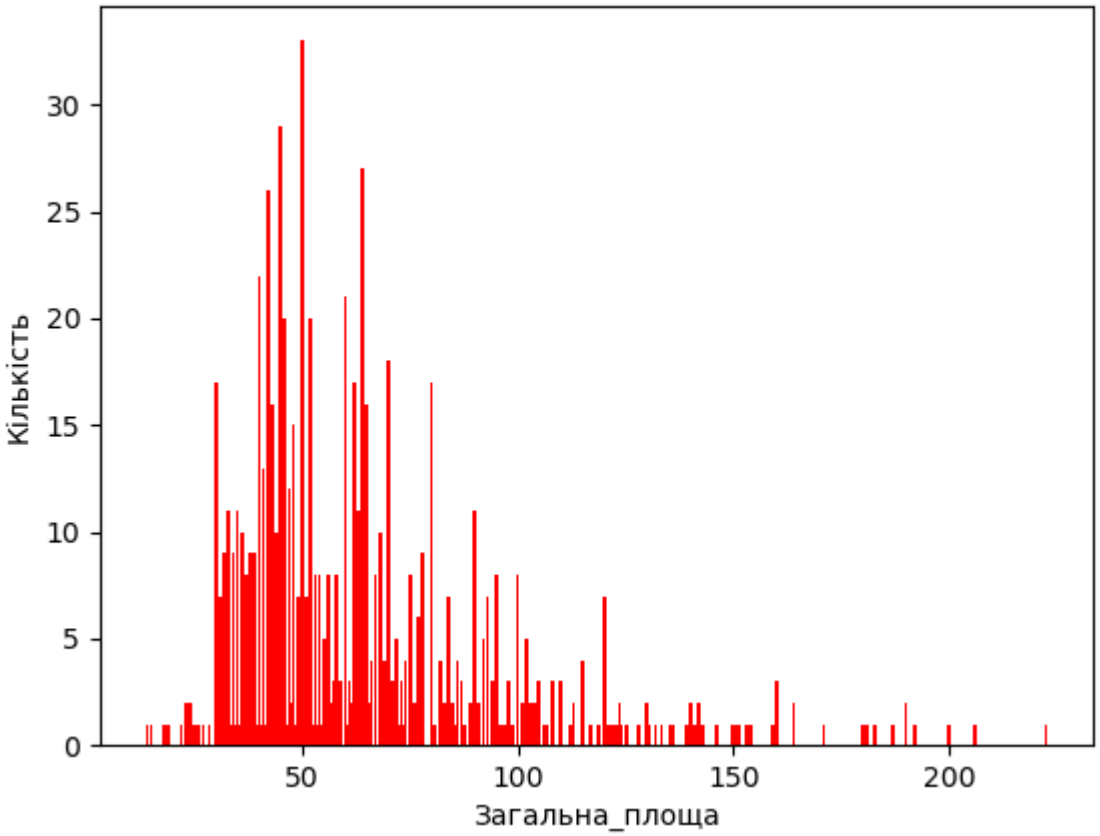

Рис. 6: Діаграма Стовпчикова "Загальна площа"

3. **Графік розподілу квартир, за загальною площею**

```
x = data['Загальна_площа']
x.plot('hist', bins=[i for i in range(0,251,25)], align='mid', color='steelblue'
plt.xlabel('Загальна_Площа')
plt.ylabel('Кількість')
plt.show()
```

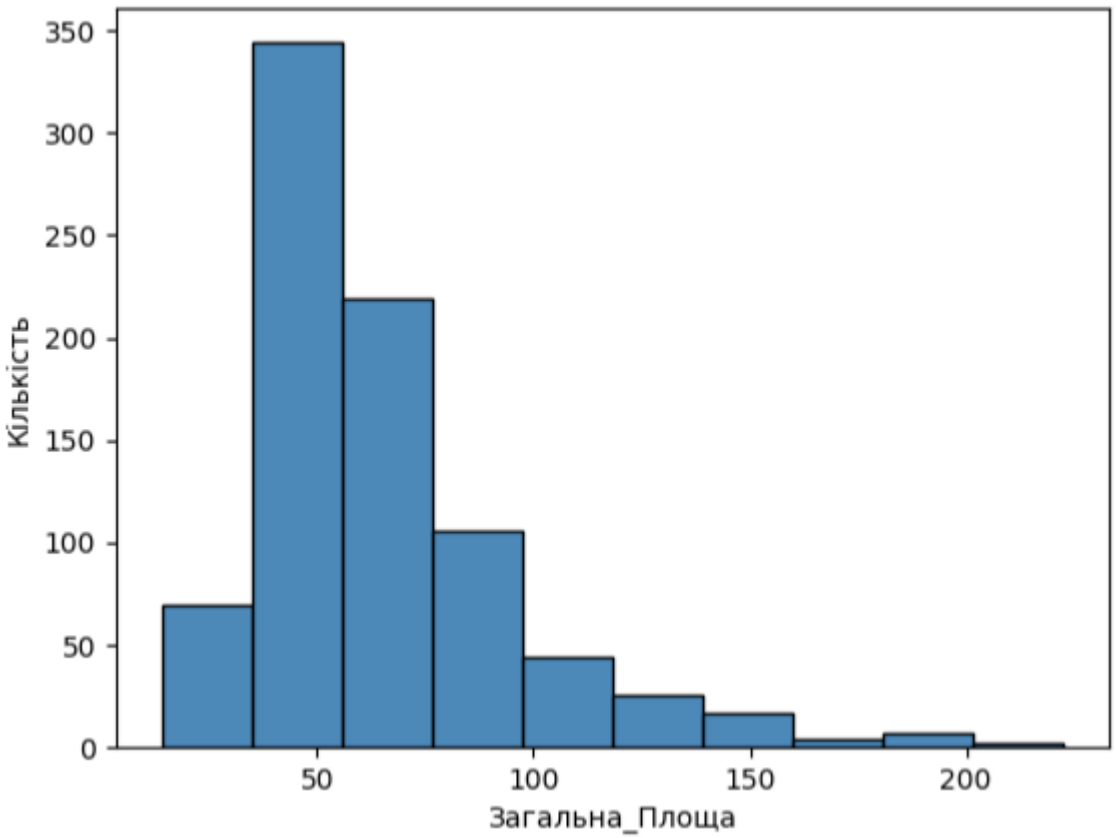

Рис. 7: Графік розподілу "Загальна площа"

4. **Графік розсіювання - залежність між ціною і загальною площею**

```
x = data['Загальна_площа']
y = data['Ціна']
plt.scatter(x, y, 3, c='g')
plt.yticks(np.arange(0, 12500001, 2500000))
plt.xticks(np.arange(0, 201, 50))
plt.xlabel('Загальна_площа')
plt.ylabel('Ціна')
plt.show()
```

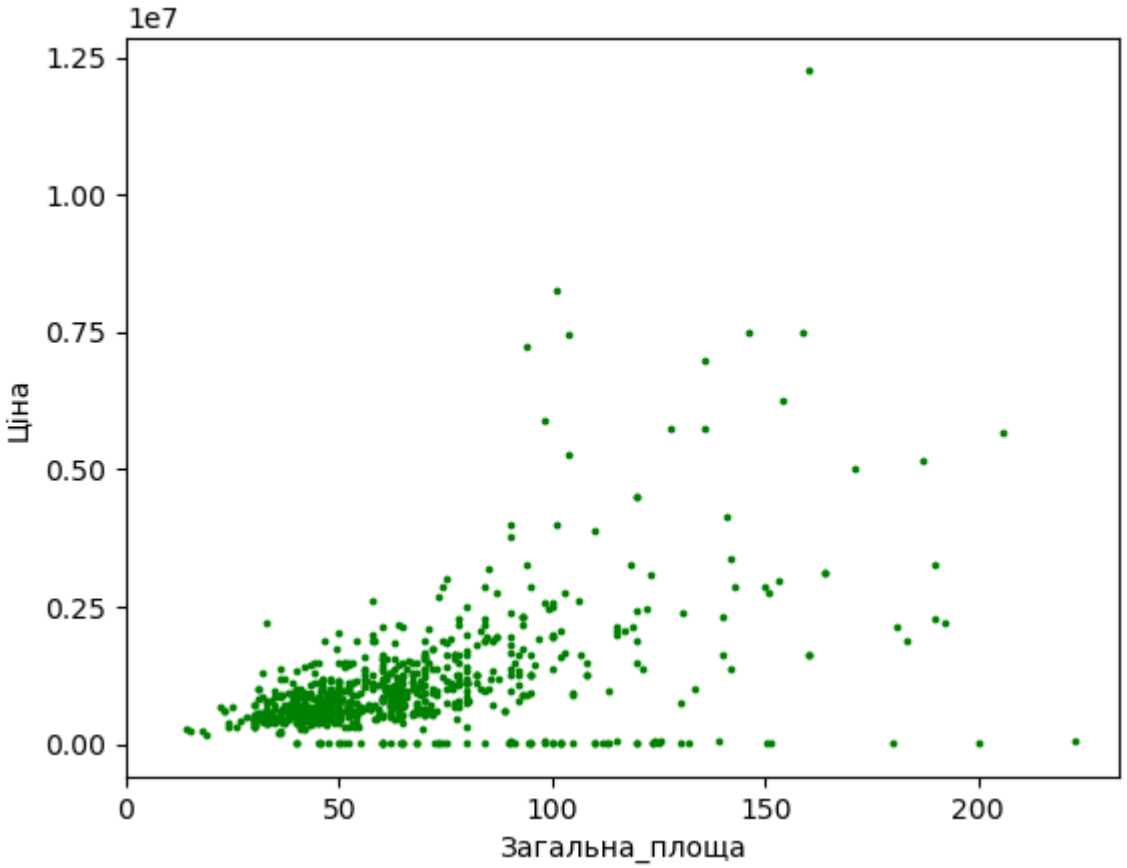

Рис. 8: Графік розсіювання "Ціна і загальна площа"

На цьому графіку окрім тренду більша площадь = більша ціна можна побачити аномальну лінію ціна 0, що певно є результатом незаповнених даних про ціну квартири на оголошеннях.

Деякі інші аномалії певно спричинені іншими залежностями ціни квартири, наприклад її розташуванням, або "видом з вікна".

5. **Графік розподілу цін по містах**

```
plt.figure(figsize=(15, 15))
ax=sns.boxplot(y='Місто', x="Ціна",orient='h', data=data, linewidth=2)
plt.show()
```

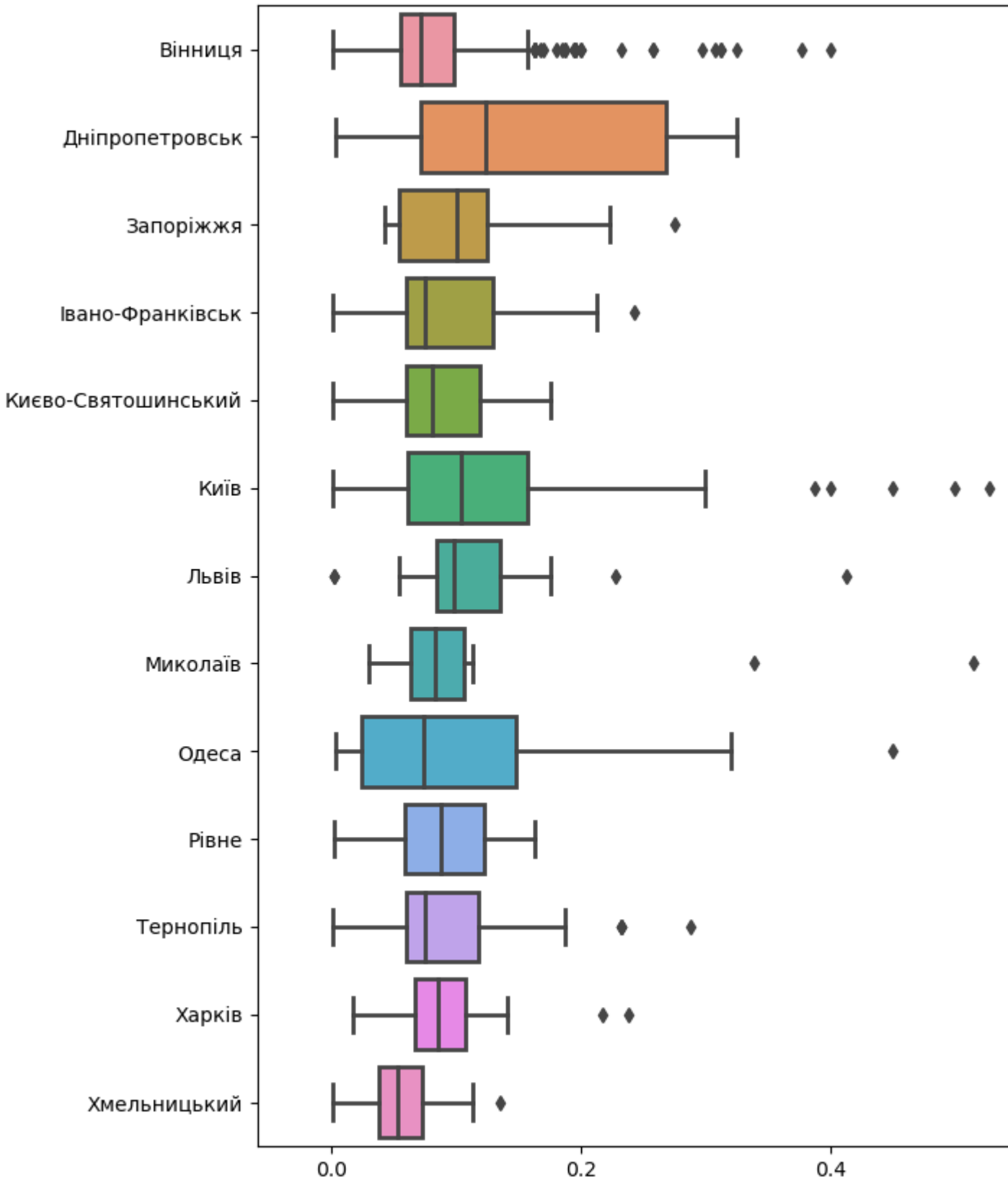

Рис. 9: Графік розподілу "Ціни по містах"

На цьому графіку можна помітити, що найбільший розкид у Одесі, Києві, та Дніпропе-

тровську. Найбільше аномалій у Києві, що певно спричинено підвищенним попитом на квартири у місті-столиці, а також більша наявність "великих"квартир у виборці(Які в свою чергу продаються дорожче).

3. Висновки

Виконуючи цю лабораторну роботу я навчився зчитувати .csv файли, будувати графіки і діаграми з отриманих даних, а також аналізувати їх.