

**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені
ТАРАСА ШЕВЧЕНКА**



ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра прикладних інформаційних систем

Звіт до лабораторної роботи №2

з курсу

«Data Science та Big Data»

Студентки 4 курсу

групи ПП-41

спеціальності 122 «Комп'ютерні науки»

ОП «Прикладне програмування»

Штось Софії Максимівни

Викладач:

Білий Р. О.

Київ – 2023

Тема роботи: Розвідувальний аналіз даних (EDA). Складання аналітичного звіту.

Мета роботи: Метою лабораторної роботи є отримання практичних навичок виконання розвідувального аналізу даних, використовуючи пакети `jupyter`, `pandas`, `seaborn`. Ознайомлення з методологією складання аналітичного звіту для зовнішнього користувача інформаційного продукту.

Контекст

Ви – щойно нанятий data analyst у великій американській компанії, яка працює на ринку нерухомості США. На черговому засіданні ваш бос дав вам завдання зробити аналітичний звіт по цікавому йому сегменту ринку - Нью-Йорку.

Завдання для виконання

- Виконайте дослідження domain experience стосовно американського ринку нерухомості. Ознайомтесь з декількома прикладами аналітичних продуктів від топових гравців на американському ринку, направлених на інвесторів. Питання, які потрібно опрацювати:

- Як топові компанії на ринку складають звіти по нерухомості?
- Які графіки використовуються для донесення інформації?
- Які співвідношення між якими даними по ринку є показовими для інвесторів / керівників агенцій нерухомості?
- Яка термінологія використовується для опису закономірностей на ринку нерухомості?

- Завантажити файли з даними у папку проекту з посилання:
<https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page>
- Очистити дані.
- Виконайте розвідувальний аналіз, щоб дізнатися, де є викиди або відсутні значення, вирішіть, як ви їх обробляти, переконайтеся, що дані відформатовані правильно, значення, які ви вважаєте числовими, розглядаються як такі і т.д.

- Виконайте аналіз розвідувальних даних (отриманих результатів) для візуалізації та зіставлення за житловими масивами та за часом. Почніть шукати осмислені закономірності у цьому наборі.
- Зберіть висновки до невеликий звіт для генерального директора (графіки, висновки з текстом у окремому файлі), який потребує належного оформлення висновків, структури тощо.

Хід роботи

- Як топові компанії на ринку складають звіти по нерухомості?

Топові компанії включають в звітах по нерухомості таку інформацію, як середня ціна житла, порівняння з минулими місяцями або роками, порівняння цін за типами житла, локацією тощо. Порівнюється ціна нерухомості також за її площею. Також, відстежують час, за який житло купується та загальну кількість нерухомості на ринку з часом, зокрема, за кварталами року.

- Які графіки використовуються для донесення інформації?

Для донесення інформації використовують такі типи графіків, як гістограми, кругові діаграми, точкові діаграми, лінійні діаграми.

- Які співвідношення між якими даними по ринку є показовими для інвесторів / керівників агенцій нерухомості?

Цінова динаміка. Інвестори слідкують за змінами вартості нерухомості на ринку. Аналіз цінової динаміки дозволяє прогнозувати тенденції та приймати рішення щодо купівлі, утримання чи продажу власності.

Географічні та демографічні аспекти. Відслідковується попит на ринку в різних регіонах, а також зміни в демографії та їх вплив на ринок.

Економічні показники. Аналіз економічного стану регіону, таких як безробіття, рівень доходів і зростання населення, може служити важливими факторами для передбачення перспектив ринку нерухомості.

- Яка термінологія використовується для опису закономірностей на ринку нерухомості?

- *YoY (Year-over-year)* – це фінансовий термін, який використовується для порівняння даних за певний період часу з відповідним періодом попереднього року.
- *LTV (Loan-to-Value)* – це відношення кредиту до вартості: Відношення суми кредиту до ринкової вартості нерухомості.
- *Cooperative (Co-op)* – це тип власності, де мешканці володіють акціями корпорації, яка володіє нерухомістю. Мешканці отримують право на проживання на основі кількості акцій, які вони мають.
- *Condominium (Condo)* – це форма власності, де кожен власник володіє окремою одиницею та має спільну власність на обштинні приміщення та зони.
- *Buyer's Market (Ринок для покупців)* – це такі умови, коли попит на нерухомість невеликий, пропозиція перевищує його. Покупці мають більший вибір та можуть очікувати знижок, а нерухомість залишається на ринку довший час.
- *Seller's Market (Ринок для продавців)*: це ситуація, де попит на нерухомість перевищує пропозицію. Це призводить до підвищення цін, швидкої продажу, та вигод для продавців, але обмежує вибір покупців.
- *Neutral Market (Нейтральний ринок)*: це рівновага між попитом і пропозицією, що призводить до стабільних цін та розумних умов як для покупців, так і для продавців.

Досліджені джерела:

1. [https://www.rockethomes.com/real-estate-trends/ny/new-york - :~:text=Median Sold Price&text=Based on all homes sold in the last 12 months.&text=Homes in New York have,per square foot was %24619.](https://www.rockethomes.com/real-estate-trends/ny/new-york-:~:text=Median Sold Price&text=Based on all homes sold in the last 12 months.&text=Homes in New York have,per square foot was %24619.)
2. <https://www.noradarealestate.com/blog/new-york-real-estate-market/>
3. <https://www.propertyshark.com/mason/market-trends/residential/nyc/manhattan>
4. <https://www.forbes.com/advisor/mortgages/real-estate/new-york-housing-market/>
5. <https://guides.nyu.edu/realestate/marketreports>

По-перше, імпортуємо необхідні пакети та створимо DataFrame з наявних датасетів:

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

In [2]: #df = pd.read_excel(r'dataset/rollingsales_bronx.xls', skiprows=4)
df = pd.concat([pd.read_excel(f'dataset/{file}', skiprows=4) for file in ["rollingsales_bronx.xls", "rollingsales_br

In [3]: df
Out[3]:
```

	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASE-MENT	BUILDING CLASS AT PRESENT	ADDRESS	APARTMENTNUMBER	...	RESIDENTIAL CC UNITS
0	2	BATHGATE	01 ONE FAMILY HOMES	1	3028	25		A5	412 EAST 179TH STREET	...		1
1	2	BATHGATE	01 ONE FAMILY HOMES	1	3039	28		A1	2329 WASHINGTON AVENUE	...		1
2	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	39		A1	2075 BATHGATE AVENUE	...		1
3	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	52		A1	2047 BATHGATE AVENUE	...		1
4	2	BATHGATE	02 TWO FAMILY HOMES	1	2900	61		S2	406 EAST TREMONT AVENUE	...		2
...
85970	5	WOODROW	02 TWO FAMILY HOMES	1	7349	10		B9	63 PHEASANT LANE	...		2
85971	5	WOODROW	02 TWO FAMILY HOMES	1	7349	35		B9	33 QUAIL LANE	...		2
85972	5	WOODROW	02 TWO FAMILY HOMES	1	7351	11		B2	40 HERRICK AVENUE	...		2

Далі видаляємо дуплікати:

```
In [4]: df = df.drop_duplicates()
df = df.reset_index(drop=True)
df
Out[4]:
```

	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASE-MENT	BUILDING CLASS AT PRESENT
0	2	BATHGATE	01 ONE FAMILY HOMES	1	3028	25		A5
1	2	BATHGATE	01 ONE FAMILY HOMES	1	3039	28		A1 W
2	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	39		A1
3	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	52		A1
4	2	BATHGATE	02 TWO FAMILY HOMES	1	2900	61		S2
...
83848	5	WOODROW	02 TWO FAMILY HOMES	1	7349	10		B9 6
83849	5	WOODROW	02 TWO FAMILY HOMES	1	7349	35		B9
83850	5	WOODROW	02 TWO FAMILY HOMES	1	7351	11		B2
83851	5	WOODROW	22 STORE BUILDINGS	4	7100	16		K6
83852	5	WOODROW	22 STORE BUILDINGS	4	7105	520		K6 2

83853 rows x 21 columns

Відформатуємо назву стовпчиків SALE PRICE, EASEMENT, APARTMENT NUMBER.

```
In [5]: df.rename(columns={'SALE PRICE': 'SALE PRICE', 'EASEMENT': 'EASEMENT', 'APARTMENT NUMBER': 'APARTMENT NUMBER'},
df
Out[5]:
```

	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASEMENT	BUILDING CLASS AT PRESENT	ADDRESS	APARTMENT NUMBER	...	RESIDENTIAL UNITS	COMMERCIAL UNITS
0	2	BATHGATE	01 ONE FAMILY HOMES	1	3028	25		A5	412 EAST 179TH STREET	...		1	
1	2	BATHGATE	01 ONE FAMILY HOMES	1	3039	28		A1	2329 WASHINGTON AVENUE	...		1	
2	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	39		A1	2075 BATHGATE AVENUE	...		1	
3	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	52		A1	2047 BATHGATE AVENUE	...		1	
4	2	BATHGATE	02 TWO FAMILY HOMES	1	2900	61		S2	406 EAST TREMONT AVENUE	...		2	
...
83848	5	WOODROW	02 TWO FAMILY HOMES	1	7349	10		B9	63 PHEASANT LANE	...		2	
83849	5	WOODROW	02 TWO FAMILY HOMES	1	7349	35		B9	33 QUAIL LANE	...		2	
83850	5	WOODROW	02 TWO FAMILY HOMES	1	7351	11		B2	40 HERRICK AVENUE	...		2	
83851	5	WOODROW	22 STORE BUILDINGS	4	7100	16		K6	639 VETERANS ROAD WEST	...		0	
83852	5	WOODROW	22 STORE BUILDINGS	4	7105	520		K6	2790 ARTHUR KILL ROAD	...		0	

83853 rows x 21 columns

Переглянемо, який вигляд має наразі датасет.

```
In [6]: df.head(60)
```

TAX CLASS AT PRESENT	BLOCK	LOT	EASEMENT	BUILDING CLASS AT PRESENT	ADDRESS	APARTMENT NUMBER	...	RESIDENTIAL UNITS	COMMERCIAL UNITS	TOTAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YEAR BUILT	TAX CLASS AT TIME OF SALE	BUILDING CLASS AT TIME OF SALE	SALE PRICE	SALE DATE
1	3028	25		A5	412 EAST 179TH STREET	...		1	0	1	1842	2048	1901	1	A5	355000	2013-07-08
1	3039	28		A1	2329 WASHINGTON AVENUE	...		1	0	1	1103	1290	1910	1	A1	474819	2013-05-20
1	3046	39		A1	2075 BATHGATE AVENUE	...		1	0	1	1986	1344	1899	1	A1	210000	2013-03-12
1	3046	52		A1	2047 BATHGATE AVENUE	...		1	0	1	2329	1431	1901	1	A1	343116	2013-07-01

Форматування адреси різниться між записами. В деяких рядках є кома, після якої вказаний номер будинку. Розділимо стовпчик адреси на назву вулиці та номер будинку.

```
In [7]: df['ADDRESS'].str.split(',', 1, expand=True)
Out[7]:
```

	0	1
0	412 EAST 179TH STREET	None
1	2329 WASHINGTON AVENUE	None
2	2075 BATHGATE AVENUE	None
3	2047 BATHGATE AVENUE	None
4	406 EAST TREMONT AVENUE	None
...
83848	63 PHEASANT LANE	None
83849	33 QUAIL LANE	None
83850	40 HERRICK AVENUE	None
83851	639 VETERANS ROAD WEST	None
83852	2790 ARTHUR KILL ROAD	None

83853 rows x 2 columns

```
In [8]: df[['STREET ADDRESS', 'BUILDING ADDRESS']] = df['ADDRESS'].str.split(',',1,expand=True)
df = df.drop(columns='ADDRESS')
df
```

Out[8]:

	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASEMENT	BUILDING CLASS AT PRESENT	APARTMENT NUMBER	ZIP CODE	...	TOTAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YE/ BUI
0	2	BATHGATE	01 ONE FAMILY HOMES	1	3028	25		A5		10457	...	1	1842	2048	19
1	2	BATHGATE	01 ONE FAMILY HOMES	1	3039	28		A1		10458	...	1	1103	1290	19
2	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	39		A1		10457	...	1	1986	1344	18
3	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	52		A1		10457	...	1	2329	1431	19
4	2	BATHGATE	02 TWO FAMILY HOMES	1	2900	61		S2		10457	...	3	1855	4452	19
...
83848	5	WOODROW	02 TWO FAMILY HOMES	1	7349	10		B9		10309	...	2	2590	2450	19
83849	5	WOODROW	02 TWO FAMILY HOMES	1	7349	35		B9		10309	...	2	2255	2377	19
83850	5	WOODROW	02 TWO FAMILY HOMES	1	7351	11		B2		10309	...	2	4000	2962	20
83851	5	WOODROW	22 STORE BUILDINGS	4	7100	16		K6		10309	...	1	21663	6950	20
83852	5	WOODROW	22 STORE BUILDINGS	4	7105	520		K6		10309	...	1	489656	159600	20

Підрахуємо, скільки нульових значень має кожен стовпчик.

```
In [9]: zero_counts_per_column = (df == 0).sum()
print(zero_counts_per_column)
```

```
BOROUGH 0
NEIGHBORHOOD 0
BUILDING CLASS CATEGORY 0
TAX CLASS AT PRESENT 0
BLOCK 0
LOT 0
EASEMENT 0
BUILDING CLASS AT PRESENT 0
APARTMENT NUMBER 0
ZIP CODE 64
RESIDENTIAL UNITS 29433
COMMERCIAL UNITS 76101
TOTAL UNITS 20314
LAND SQUARE FEET 39518
GROSS SQUARE FEET 41880
YEAR BUILT 11113
TAX CLASS AT TIME OF SALE 0
BUILDING CLASS AT TIME OF SALE 0
SALE PRICE 27489
SALE DATE 0
STREET ADDRESS 0
BUILDING ADDRESS 0
dtype: int64
```

Звернемо увагу на те, що значення 'ZIP CODE', 'RESIDENTIAL UNITS', 'COMMERCIAL UNITS', 'TOTAL UNITS', 'LAND SQUARE FEET', 'GROSS SQUARE FEET', 'YEAR BUILT', 'SALE PRICE' дорівнюють 0 в деяких рядках. Оскільки доволі багато рядків мають нулі, ми їх так і залишимо, але ці нулі потрібно ігнорувати під час обчислення будь-якого статистичного значення або побудови графіка розподілу.

Для 'ZIP CODE' заповнимо два нульових значення значеннями відповідних попередніх рядків та перевіримо формати значень.

```
In [10]: df['ZIP CODE'].replace(0, np.nan, inplace=True)
df['ZIP CODE'].fillna(method='ffill', inplace=True)
df['ZIP CODE'] = df['ZIP CODE'].astype('int64')
```

Перевіримо формати значень:

```
In [11]: print(df.dtypes)
```

```
BOROUGH                int64
NEIGHBORHOOD           object
BUILDING CLASS CATEGORY object
TAX CLASS AT PRESENT   object
BLOCK                 int64
LOT                  int64
EASEMENT              object
BUILDING CLASS AT PRESENT object
APARTMENT NUMBER       object
ZIP CODE              int64
RESIDENTIAL UNITS      int64
COMMERCIAL UNITS       int64
TOTAL UNITS            int64
LAND SQUARE FEET      int64
GROSS SQUARE FEET     int64
YEAR BUILT             int64
TAX CLASS AT TIME OF SALE object
BUILDING CLASS AT TIME OF SALE object
SALE PRICE             int64
SALE DATE              datetime64[ns]
STREET ADDRESS         object
BUILDING ADDRESS       object
dtype: object
```

	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASEMENT	BUILDING CLASS AT PRESENT	APARTMENT NUMBER	ZIP CODE	...	TOTAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YEAR BUILT
0	2	BATHGATE	01 ONE FAMILY HOMES	1	3028	25		A5		10457	...	1	1842	2048	19
1	2	BATHGATE	01 ONE FAMILY HOMES	1	3039	28		A1		10458	...	1	1103	1290	19
2	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	39		A1		10457	...	1	1986	1344	18
3	2	BATHGATE	01 ONE FAMILY HOMES	1	3046	52		A1		10457	...	1	2329	1431	19
4	2	BATHGATE	02 TWO FAMILY HOMES	1	2900	61		S2		10457	...	3	1855	4452	19
...
83848	5	WOODROW	02 TWO FAMILY HOMES	1	7349	10		B9		10309	...	2	2590	2450	19
83849	5	WOODROW	02 TWO FAMILY HOMES	1	7349	35		B9		10309	...	2	2255	2377	19
83850	5	WOODROW	02 TWO FAMILY HOMES	1	7351	11		B2		10309	...	2	4000	2962	20
83851	5	WOODROW	22 STORE BUILDINGS	4	7100	16		K6		10309	...	1	21663	6950	20
83852	5	WOODROW	22 STORE BUILDINGS	4	7105	520		K6		10309	...	1	489656	159600	20

83853 rows x 22 columns

Бачимо, що дані відформатовані належним чином, а отже можна приступати до візуалізації, яка описана у файлі DSBD_Звіт_Лаб2_ПП41_ШтоньС.

Висновок: Отже, у ході цієї лабораторної роботи було отримано навички проведення розвідувального аналізу даних (EDA) та складання аналітичного звіту.