КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

імені ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра прикладних інформаційних систем

Звіт

до виконання лабораторної роботи №2

з дисципліни « **Data Science та Big Data** »

на тему:

« Розвідувальний аналіз даних (EDA). Складання аналітичного звіту »

Виконано:

студ. групи ПП- 41, підгрупа 2

Шкандюк Анною Леонідівною

Перевірено:

к.т.н., доц. Білий Р.О.

Київ – 2023

1. **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.

2. **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.

На базі цього виконати наступні завдання:

3. **Хід виконання:**

1. 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:

a. Як топові компанії на ринку складають звіти по нерухомості?

Топові компанії на ринку нерухомості складають звіти, які містять такі основні розділи:

- **Аналіз ринку:** Загальний огляд стану ринку, включаючи зміни цін, обсяг транзакцій, тенденції та прогнози.
- **Фінансовий звіт:** Подання фінансових показників, таких як прибуток, витрати, прибутковість проектів та інші фінансові метрики.
- **Демографічні дані:** Аналіз цільової аудиторії та демографічних характеристик регіону.

- **Правові аспекти:** Врахування регулятивних змін, законодавства та інших юридичних аспектів, які можуть вплинути на ринок.

b. Які графіки використовуються для донесення інформації?
   - **Графіки цін:** Лінійні графіки, стовпчасті гістограми та кругові діаграми для візуалізації змін цін на ринку.
   - **Карти ринку:** Географічні карти для відображення розташування нерухомості та динаміки цін за регіонами.
   - **Графіки обсягу транзакцій:** Графіки обсягу продажів та кількість угод для визначення активності ринку.
   - **Порівняльні графіки:** Графіки порівняння різних показників, таких як ціни на ринку в різних регіонах чи типах нерухомості.

c. Які співвідношення між якими даними по ринку є показовими для інвесторів / керівників агенцій нерухомості?
   - **Ціна до оренди:** Співвідношення ціни купівлі до потенційного орендного доходу важливо для інвесторів.
   - **Обсяг транзакцій та ліквідність:** Визначення активності ринку та ліквідності майна.
   - **Динаміка цін та прибутковість:** Розуміння тенденцій у зміні цін та прибутковості інвестицій.
   - **Демографічний аналіз:** Врахування демографічних факторів для визначення цільової аудиторії.

d. Яка термінологія використовується для опису закономірностей на ринку нерухомості?
   - **Cap Rate (Capitalization Rate):** Співвідношення між очікуваним річним чистим прибутком та вартістю нерухомості.
   - **ROI (Return on Investment):** Показник прибутковості інвестицій.
   - **Absorption Rate:** Швидкість, з якою нерухомість продається на ринку.
   - **Listings vs. Sales:** Відсоток оголошень, які фактично продаються.

- **Days on Market (DOM):** Кількість днів, які нерухомість перебуває на ринку перед продажем.

2. Завантажити файли з даними у папку проекту з посилання:

https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

3. Очистити дані.

4. Виконайте розвідувальний аналіз, щоб дізнатися, де є викиди або відсутні значення, вирішіть, як ви їх будете обробляти, переконайтеся, що дати відформатовані правильно, значення, які ви вважаєте числовими, розглядаються як такі і т.д.

```
       BOROUGH         BLOCK           LOT  EASEMENT       ZIP CODE   \
count  73305.000000  73305.000000  73305.000000       0.0  73300.000000
mean       2.961490   4172.508205    389.437146       NaN  10848.080246
std        1.298879   3548.291483    659.348616       NaN    566.823256
min        1.000000      1.000000      1.000000       NaN  10001.000000
25%        2.000000   1263.000000     22.000000       NaN  10304.000000
50%        3.000000   3238.000000     51.000000       NaN  11209.000000
75%        4.000000   6201.000000   1003.000000       NaN  11356.000000
max        5.000000  16350.000000   9079.000000       NaN  11697.000000

       RESIDENTIAL UNITS  COMMERCIAL UNITS  TOTAL UNITS  LAND SQUARE FEET  \
count       55472.000000      42268.000000  57801.000000      3.993900e+04
mean            3.565024          0.413149      3.723500      7.504976e+03
```

```
std              24.601775           8.322000      25.200924       1.395299e+05
min               0.000000           0.000000       0.000000       0.000000e+00
25%               1.000000           0.000000       1.000000       2.000000e+03
50%               1.000000           0.000000       1.000000       2.500000e+03
75%               2.000000           0.000000       2.000000       4.000000e+03
max            1019.000000        1178.000000    1178.000000       9.166500e+06

       GROSS SQUARE FEET      YEAR BUILT   TAX CLASS AT TIME OF SALE   \
count           3.993900e+04   68430.000000                          73305.000000
mean            6.812155e+03    1952.779906                              1.657977
std             4.865590e+04       36.479919                             0.795865
min             0.000000e+00      220.000000                             1.000000
25%             1.381500e+03    1925.000000                              1.000000
50%             1.988000e+03    1950.000000                              2.000000
75%             2.848000e+03    1975.000000                              2.000000
max             2.400000e+06    2023.000000                              4.000000

         SALE                                                          PRICE
count                                                          7.330500e+04
mean                                                           1.243891e+06
std                                                            6.348467e+06
min                                                            0.000000e+00
25%                                                            0.000000e+00
50%                                                            5.100000e+05
75%                                                            9.990000e+05
max     5.981558e+08
```

Перші три рядки вказують на кількість, середнє значення та стандартне відхилення для різних характеристик:

1. `count`: кількість непорожніх значень для кожного стовпця.

2. `mean`: середнє значення для кожного стовпця.

3. `std`: стандартне відхилення для кожного стовпця.

Далі йдуть статистичні показники для кількісних ознак:

4. `min`: мінімальне значення для кожного стовпця.

5. `25%`: 25-й перцентиль (перший квартиль).

6. `50%`: 50-й перцентиль (медіана).

7. `75%`: 75-й перцентиль (третій квартиль).

8. `max`: максимальне значення для кожного стовпця.

BUILDING CLASS CATEGORY	
25 LUXURY HOTELS	3.997709e+07
39 TRANSPORTATION FACILITIES	3.955248e+07
33 EDUCATIONAL FACILITIES	1.712481e+07
26 OTHER HOTELS	1.450344e+07
45 CONDO HOTELS	1.281863e+07
21 OFFICE BUILDINGS	1.226184e+07
08 RENTALS - ELEVATOR APARTMENTS	1.104416e+07
38 ASYLUMS AND HOMES	8.863923e+06
32 HOSPITAL AND HEALTH FACILITIES	7.849279e+06
31 COMMERCIAL VACANT LAND	7.213012e+06
34 THEATRES	6.473457e+06
46 CONDO STORE BUILDINGS	6.047294e+06
27 FACTORIES	5.380449e+06
41 TAX CLASS 4 - OTHER	4.862317e+06
29 COMMERCIAL GARAGES	4.722574e+06
28 COMMERCIAL CONDOS	4.226977e+06
30 WAREHOUSES	4.153243e+06
11 SPECIAL CONDO BILLING LOTS	3.642315e+06
22 STORE BUILDINGS	3.168123e+06
43 CONDO OFFICE BUILDINGS	2.259563e+06
13 CONDOS - ELEVATOR APARTMENTS	1.921218e+06
14 RENTALS - 4-10 UNIT	1.918221e+06
07 RENTALS - WALKUP APARTMENTS	1.735802e+06
36 OUTDOOR RECREATIONAL FACILITIES	1.645407e+06
47 CONDO NON-BUSINESS STORAGE	1.585135e+06
37 RELIGIOUS FACILITIES	1.564733e+06
44 CONDO PARKING	1.401669e+06
15 CONDOS - 2-10 UNIT RESIDENTIAL	1.356944e+06

```
17 CONDO COOPS                                1.176343e+06
04 TAX CLASS 1 CONDOS                         1.152435e+06
16 CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT    9.433807e+05
12 CONDOS - WALKUP APARTMENTS                 9.374980e+05
49 CONDO WAREHOUSES/FACTORY/INDUS             8.284740e+05
10 COOPS - ELEVATOR APARTMENTS                7.964560e+05
48 CONDO TERRACES/GARDENS/CABANAS             7.710765e+05
35 INDOOR PUBLIC AND CULTURAL FACILITIES      7.143043e+05
05 TAX CLASS 1 VACANT LAND                    7.102914e+05
09 COOPS - WALKUP APARTMENTS                  6.249592e+05
03 THREE FAMILY DWELLINGS                     6.022830e+05
02 TWO FAMILY DWELLINGS                       5.796386e+05
01 ONE FAMILY DWELLINGS                       5.466043e+05
42 CONDO CULTURAL/MEDICAL/EDUCATIONAL/ETC     5.320909e+05
06 TAX CLASS 1 - OTHER                        4.956698e+05
40 SELECTED GOVERNMENTAL FACILITIES           0.000000e+00
```

Це статистичні показники для категоріальної ознаки "BUILDING CLASS CATEGORY". Кожен рядок виводу показує середнє значення ціноутворення для конкретної категорії. Наприклад, середня ціна для "25 LUXURY HOTELS" дорівнює `3.997709e+07`.

Як бачимо, відсутні значення автоматично заміняються на Nan, що дає нам змогу не враховувати їх при підрахунку статистики. Усі числові значення розглядаються як такові та можуть бути використані при підрахунках.

5. Виконайте аналіз розвідувальних даних (отриманих результатів) для візуалізації та зіставлення за житловими масивами та за часом. Почніть шукати осмислені закономірності у цьому наборі.

Огляд Даних:

Борги: Датасет містить дані для всіх п'яти боро Нью-Йорка (Борги 1-5).

Загальна кількість одиниць: Більшість об'єктів в датасеті, схоже, мають відносно низьку кількість одиниць.

Ціна продажу: Середня ціна продажу становить приблизно 1,67 мільйона доларів, існує широкий діапазон цін за стандартним відхиленням.

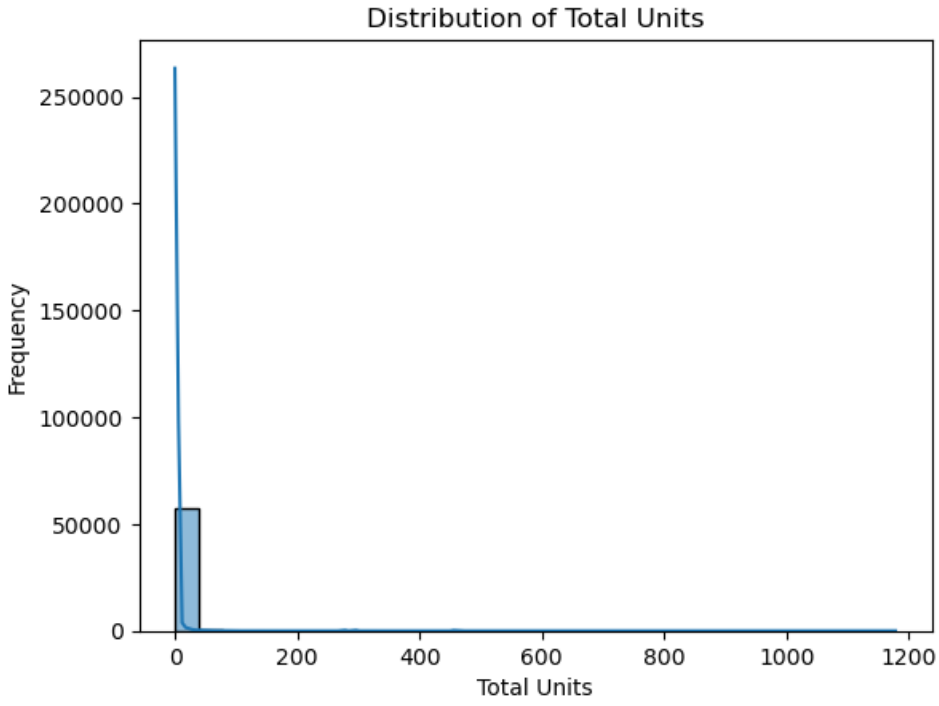
Середня Ціна за Типом Власності:

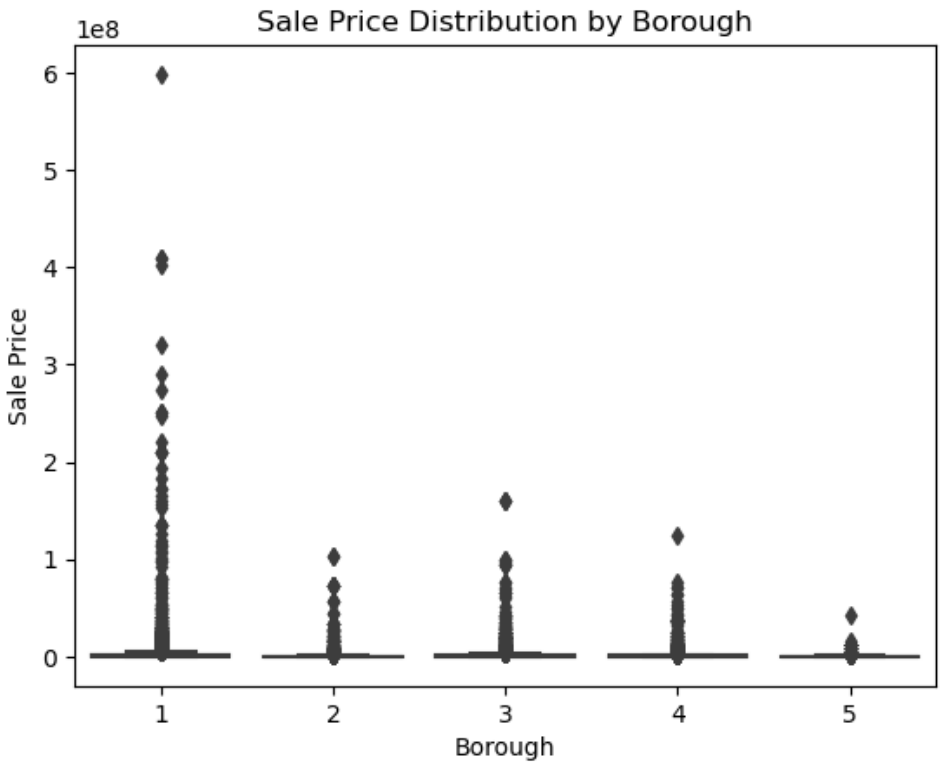
Різні категорії типів будівель мають відмінні середні ціни продаж.

Категорії, такі як 'ОДНА СІМ' та 'ДВІ СІМ' схильні мати більш низькі середні ціни продажу.

Категорії, такі як 'КООПЕРАТИВИ - КВАРТИРИ З ЛІФТОМ' та 'КОНДОМІНІУМИ - БАГАТОКВАРТИРНІ БУДІВЛІ', мають високі середні ціни продажу.

Візуалізації:





Розподіл Кількості Одиниць: Гістограма показує, що більшість об'єктів мають відносно низьку кількість одиниць.

Розподіл Цін на Продаж по Областям: Ящиковий графік вказує на варіації цін продажу в різних областях, надаючи порівняльний огляд.

Додаткові Розгляди:

Аномалії в Даних: Мінімальне значення 'РІК ПОБУДОВИ' здається надто низьким (220), що може свідчити про потенційну проблему або аномалію в даних."

```
# Display unique values in the 'YEAR BUILT' column
uniqueYears = data['YEAR BUILT'].unique()
print("Unique Years:", uniqueYears)
```

```
Unique Years: [1900. 1910. 1904. 1913. 1930. 1925. 1920. 1928. 1923.
2001. 1950. 1940.
 1937. 1929.   nan 2014. 2005. 2008. 2009. 1960. 2017. 2021. 2000. 1958.
 1865. 1901. 1905. 1885. 1850. 1938. 1944. 1974. 2002. 1864. 1921. 1917.
 1911. 2003. 1983. 1963. 1926. 1902. 1889. 1898. 1939. 1918. 1927. 1909.
 2015. 2007. 2018. 2013. 2020. 2019. 2016. 2006. 1989. 1985. 1984. 1912.
```

```
 2004. 1899. 1987. 1875. 2012. 1973. 2011. 1922. 1851. 1932. 1931. 1980.
 1908. 1919. 1914. 1880. 1915. 1990. 1907. 1991. 1965. 1890. 2010. 1924.
 1860. 1975. 1896. 1957. 1986. 1988. 1946. 1888. 1870. 1906. 1956. 1982.
 1903. 1967. 1840. 1969. 1968. 1964. 1955. 1961. 1947. 1945. 1959. 1962.
 1951. 1972. 1976. 1952. 1948. 1941. 1895. 1966. 1981. 1954. 1841. 1839.
 1842. 1869. 1836. 1999. 1997. 1994. 2022. 2023. 1892. 1992. 1916. 1934.
 1949. 1935. 1953. 1998. 1979. 1977. 1970. 1942. 1933. 1978. 1800. 1871.
 1879. 1886. 1872. 1971. 1830. 1936. 1996. 1884. 1993. 1995. 1897. 1893.
  220. 1858. 1834. 1829. 1843. 1835. 1849. 1848. 1846. 1844. 1831. 1845.
 1855. 1832. 1867. 1883. 1887. 1882. 1877. 1802. 1881. 1891. 1943. 1878.]
```

Дійсно маємо аномалію в році 220. Виведемо той рядок:

```
print(data.loc[data['YEAR BUILT'] == 220])
```

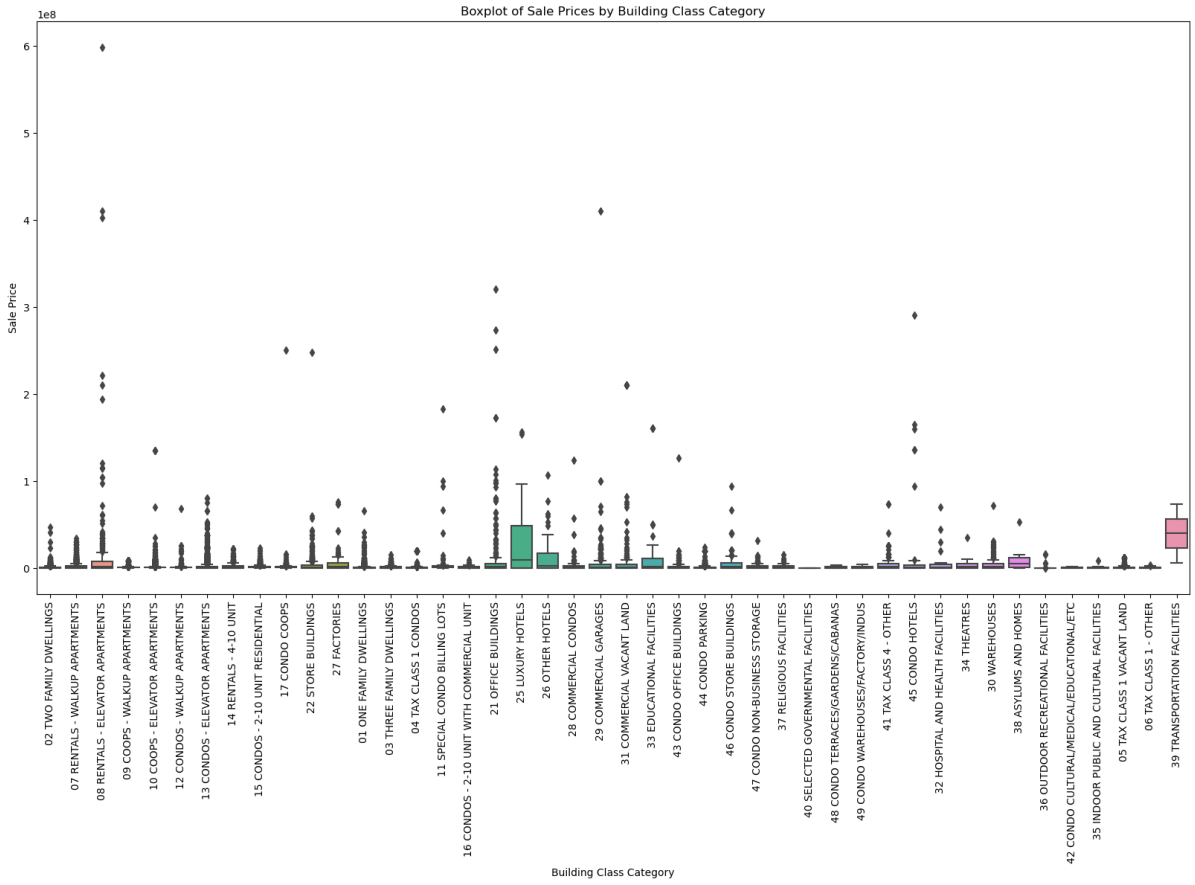
```
      BOROUGH NEIGHBORHOOD          BUILDING CLASS CATEGORY  \
1037        3    BAY RIDGE  13 CONDOS - ELEVATOR APARTMENTS
     TAX CLASS AT PRESENT  BLOCK   LOT  EASEMENT BUILDING CLASS AT PRESENT  \
1037                    2   6133  1207       NaN                         R4
                    ADDRESS APARTMENT NUMBER  ...  RESIDENTIAL UNITS  \
1037  9956 THIRD AVENUE, 301             301  ...                1.0
      COMMERCIAL UNITS  TOTAL UNITS  LAND SQUARE FEET  GROSS SQUARE FEET  \
1037               NaN          1.0               NaN                NaN

      YEAR BUILT  TAX CLASS AT TIME OF SALE  BUILDING CLASS AT TIME OF SALE  \
1037       220.0                          2                               R4
     SALE PRICE  SALE DATE
1037     890228 2022-11-22

[1 rows x 21 columns]
```

```
plt.figure(figsize=(20, 10))
sns.boxplot(x='BUILDING CLASS CATEGORY', y='SALE PRICE', data=data)

plt.title('Boxplot of Sale Prices by Building Class Category')
plt.xlabel('Building Class Category')
plt.ylabel('Sale Price')
plt.xticks(rotation=90, ha='center')  # Rotate x-axis labels for better readability
plt.show()
```




```
median_prices = data.groupby('BUILDING CLASS CATEGORY')['SALE PRICE'].median().sort_values(ascending=False)

# Print the median prices in a tabular format
print("Building Class Category\tMedian Sale Price")
print("-------------------------------------------------")
for category, median_price in median_prices.items():
    print(f"{category}\t\t\t${median_price:.2f}")
```

```
Building Class Category  Median Sale Price
-------------------------------------------------
39 TRANSPORTATION FACILITIES                $39552481.50
25 LUXURY HOTELS                $9375000.00
38 ASYLUMS AND HOMES                    $4947000.00
26 OTHER HOTELS                 $2600000.00
27 FACTORIES            $1888000.00
46 CONDO STORE BUILDINGS            $1800000.00
33 EDUCATIONAL FACILITIES                   $1732625.00
08 RENTALS - ELEVATOR APARTMENTS                    $1560000.00
34 THEATRES             $1477500.00
30 WAREHOUSES                   $1425000.00
21 OFFICE BUILDINGS                 $1400000.00
11 SPECIAL CONDO BILLING LOTS               $1310084.00
```

```
41 TAX CLASS 4 - OTHER              $1150000.00
14 RENTALS - 4-10 UNIT              $900000.00
47 CONDO NON-BUSINESS STORAGE             $894024.00
22 STORE BUILDINGS            $875000.00
15 CONDOS - 2-10 UNIT RESIDENTIAL               $860000.00
13 CONDOS - ELEVATOR APARTMENTS                 $850000.00
42 CONDO CULTURAL/MEDICAL/EDUCATIONAL/ETC             $838000.00
29 COMMERCIAL GARAGES               $800000.00
32 HOSPITAL AND HEALTH FACILITIES               $754713.00
43 CONDO OFFICE BUILDINGS                 $747500.00
37 RELIGIOUS FACILITIES             $700000.00
17 CONDO COOPS                $675000.00
16 CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT            $595000.00
44 CONDO PARKING              $544000.00
12 CONDOS - WALKUP APARTMENTS             $515000.00
01 ONE FAMILY DWELLINGS             $499997.00
28 COMMERCIAL CONDOS                $470000.00
04 TAX CLASS 1 CONDOS               $420000.00
10 COOPS - ELEVATOR APARTMENTS            $399000.00
07 RENTALS - WALKUP APARTMENTS            $397499.50
45 CONDO HOTELS               $365050.00
09 COOPS - WALKUP APARTMENTS              $325000.00
02 TWO FAMILY DWELLINGS             $310000.00
31 COMMERCIAL VACANT LAND                 $287500.00
06 TAX CLASS 1 - OTHER              $147500.00
48 CONDO TERRACES/GARDENS/CABANAS               $125000.00
05 TAX CLASS 1 VACANT LAND                $75000.00
49 CONDO WAREHOUSES/FACTORY/INDUS                 $45000.00
36 OUTDOOR RECREATIONAL FACILITIES              $30.00
35 INDOOR PUBLIC AND CULTURAL FACILITIES              $0.00
40 SELECTED GOVERNMENTAL FACILITIES             $0.00
03 THREE FAMILY DWELLINGS                 $0.00
```

Спостереження та можливих висновки:

Велика різниця в цінах:

- Ціни на нерухомість в різних категоріях будівель значно відрізняються, що може бути визначальним фактором при визначенні стратегії інвестування.

Високі ціни в галузі транспорту та готелів:

- Категорії, пов'язані з транспортними засобами та готелями (39, 25, 46), мають найвищі медіанні ціни. Це може вказувати на високий рівень прибутковості в цих сегментах.

Низькі ціни в галузі розваг та нерухомості для одного сімейного будинку:

- Категорії, пов'язані з розвагами (36, 35, 40), а також одноквартирні будинки (01, 03, 02), мають найнижчі медіанні ціни. Це може вказувати на менший рівень прибутковості в цих сегментах.

Широкий розмах цін:

- Є категорії з великим розмахом цін (наприклад, 39 та 25), де можуть існувати як високоприбуткові об'єкти, так і ті, що не такі прибуткові.

Популярність житлових будинків:

- Односімейні будинки (01) та будинки з двома квартирами (02) мають середні медіанні ціни, що може свідчити про їхню популярність серед покупців.

Урахування інших факторів:

- Потрібно також враховувати інші фактори, такі як розташування, стан будівлі, попит та пропозиція на ринку, щоб зробити повніший аналіз та прийняти інформовані рішення.

Ціни рівні нулю:

- Деякі категорії, такі як "INDOOR PUBLIC AND CULTURAL FACILITIES" та "OUTDOOR RECREATIONAL FACILITIES", мають

нульові медіанні ціни продажу, що може вказувати на некомерційні або не-житлові властивості.

Можливі викиди:

- Категорії з надто низькими медіанними цінами, такі як "SELECTED GOVERNMENTAL FACILITIES" та "OUTDOOR RECREATIONAL FACILITIES", можуть вимагати додаткового вивчення для виявлення можливих викидів чи унікальних характеристик.

6. Зберіть висновки у невеликий звіт для генерального директора (графіки, висновки з текстом у окремому файлі), який потребує належного оформлення висновків, структури тощо.

4. **Висновки до роботи**

Під час виконання лабораторної роботи 1, я отримала практичні навички у роботі з raw data, використовуючи пакети jupyter, pandas, seaborn.