Propreturns Data Science Internship Assignment

Web Scraping

By:- Malhar Jadhav

To-Do

Scrape the following Delhi Government website for Real Estate data using Selenium - https://esearch.delhigovt.nic.in/Complete_search.aspx

You may use the following inputs -

SRO: Central -Asaf Ali (SR III)
Locality: All localities</h7>
Registration Year: 2021-2022</h7>

Importing the required Libraries

In [1]:
```python
from selenium import webdriver
from selenium.webdriver.support.ui import Select
from selenium.webdriver.common.by import By
from time import sleep

driver = webdriver.Chrome(executable_path=r"D:/Malhar H Jadhav/Devlopment/chromedriver.exe")
driver.get("http://esearch.delhigovt.nic.in/Complete_search.aspx")

Select(driver.find_element( By.ID, "ctl00_ContentPlaceHolder1_ddl_sro_s")).select_by_visible_text("Central -Asaf Ali (SR III)")
sleep(2)

location_select = Select(driver.find_element(By.ID, "ctl00_ContentPlaceHolder1_ddl_loc_s"))
locations = [location.text for location in location_select.options if location.text != "-Select All-"]
del locations[0]

driver.quit()
```

```
C:\Users\harsh\AppData\Local\Temp\ipykernel_17576\1526894683.py:6: DeprecationWarning: executable_path has been deprecated, please pass in a Service object
  driver = webdriver.Chrome(executable_path=r"D:/Malhar H Jadhav/Devlopment/chromedriver.exe")
```

Observation: Above code will help us to get all locations present in Central Asaf Ali (SR III)

In [2]:
```python
locations
```

Out[2]:
```
['Abul Fazal Enclave*',
 'Adarsh Nagar*',
 'Ahata kidara*',
 'Ajmal Khan Road',
 'Ajmeri Gate',
 'Alipur Road*',
 'Alipur*',
 'Amarpuri*',
 'Amrit Kaur Puri',
 'Anand Niketan*',
 'Andrews Ganj*',
 'Anna Nagar (Minto Road)',
 'Anoop Nagar*',
 'Ansari Nagar*',
 'Arakarshana Road',
 'Aram Bagh',
 'Aram Nagar',
 'Arjun Nagar*',
 'Arya Nagar (Pahar Ganj)',
 'Arya Nagar*',
 'Arya Samaj Road',
 'Asaf Ali Road',
 'Ashok Nagar*',
 'Ashok Vihar*',
 'Ashoka Pahari (Manakpura )',
 'Ashram*',
 'Asola*',
 'Badarpur*',
 'Baggichi Madhodas (Jama Masjid)',
 'Bagh Raoji Colony (Manakpura)',
 'Bagh Raoji*',
 'Bagichi Allauddin (Qadam Shariff)',
 'Bagichi Allauddin*',
 'Bahadur Shah Zafar Marg',
 'Bali Nagar*',
 'Baljeet Nagar*',
 'Balmiki Basti (Minto Road)',
 'Balmiki Colony (Dev Nagar)',
 'Bangali Market*',
 'Bapa Nagar',
 'Bara Hindu Rao*',
 'Baradari (Ballimaran)',
 'Basai Darapur*',
 'Basant Nagar*',
 'Basti Harphool Singh*',
 'Basti Julahan*',
 'Batla House*',
 'Bazar Lal Kuan*',
 'Bazar Sita Ram',
 'Beadon Pura',
 'Bela Road*',
 'Beri Wala Bagh*',
 'Bhagirath Palace*',
 'Bharat Nagar*',
 'Bharthal*',
 'Bhikaji Cama Place*',
 'Bhim Nagar*',
 'Bhogal*',
 'Bijwasan*',
 'Bindapur*',
 'Birla Lines*',
 'Bulbuli Khan Darya Ganj',
 'Chamelian Road*',
 'Chanakya Puri*',
 'Chandiwalan*',
 'Chandni Mahal',
 'Chandrawal Road*',
 'Chatta Lal Mian*',
 'Chatta Lal Miya (Darya Ganj)',
 'Chawri Bazar*',
 'Chhatta Lal Mian*',
 'Chhawla*',
 'Chirag Delhi*',
 'Chitla Gate Area (Darya Ganj)',
 'Chitli Qabar*',
 'Chuna Mandi',
 'Churiwalan*',
 'Civil Lines*',
 'Dakshinpuri Extension*',
 'Dariba Kalan*',
 'Darya Ganj*',
 'Daryaganj*',
 'DDU Marg*',
 'Defence Colony*',
 'Delhi Gate Bazar',
 'Delhi Gate*',
 'Dev Nagar',
 'Dharam Pura, Chandni Chowk',
 'Dharampura*',
 'Dhaula Kuan*',
 'Dilshad Garden*',
 'Dori Walan',
 'Doriwalan*',
 'Dwarka*',
 'East of Kailash*',
 'East Patel Nagar*',
 'Faiz Bazar*',
 'Faiz Road*',
 'Farash Khana',
 'Fateh Nagar*',
 'Fatehpur Beri*',
 'Feroz Shah Kotla',
 'G.B. Road',
 'Gali Garhiya',
 'Gali Imam Wali',
 'Gali Madarsa Abdul Aziz',
 'Gali Madarsa Hussain Bux',
 'Gali Masjid Lal',
 'Gali Matia Mahal',
 'Gali Shahtara',
 'Gandhi Market Area Minto Road)',
 'Gandhi Nagar*',
 'Ganeshpura*',
 'Ganj Mir Khan*',
 'Gau Shala (Manakpura)',
 'Gaushala Marg*',
 'Gautam Nagar*',
 'Ghanta Ghar*',
 'Gopal Nagar*',
 'Govind Nagar',
 'Greater Kailash*',
 'Green Park Extension*',
 'Green Park Main*',
 'Green Park Market*',
 'Gulabi Bagh*',
 'Gulmohar Enclave*',
 'Gulmohar Park*',
 'Gurgaon Road*',
 'Hardev Puri*',
 'Hardhyan Singh Road',
 'Hari Nagar Ashram*',
 'Hari Nagar*',
 'Hauz Khas Enclave*',
 'Hauz Khas*',
 'Hauz Qazi*',
 'Hauz Quazi',
 'Hauz Rani*',
 'Haveli Azam Khan',
 'Haveli Hissamuddin Haider*',
 'Idgah Road*',
 'Inderpuri*',
 'Indira Nagar*',
 'Indra Park*',
 'Indraprasth Estate (Minto Road)',
 'Issapur*',
 'Jagatpur*',
 'Jama Masjid',
 'Jamia Nagar*',
 'Janak Park*',
 'Janak Puri*',
 'Janakpuri*',
 'Jangpura Lane*',
 'Jangpura Mathura Road*',
 'Jasola Village*',
 'Jhandewalan',
 'Jhandewalan Extn.(Manakpura)',
 'Jhandewalan Road*',
 'Jia Sarai*',
 'Joga Bai*',
 'Jogabai*',
 'Jogiwara*',
 'Joshi Road',
 'Kailash Colony*',
 'Kala Mahal',
 'Kala Masjid',
 'Kalinidi Colony*',
 'Kalka Ji*',
 'Kalkaji*',
 'Kamla Market',
 'Kapashera*',
 'Karol Bagh',
 'Kaseruwalan*',
 'Kashmere Gate*',
 'Katra Chhotey Lal (Darya Ganj',
 'Kaushri Wallan',
 'Kautilya Marg*',
 'Khaira*',
 'Khalsa Nagar',
 'Khari Baoli*',
 'Khirki Extension*',
 'Kilokari*',
 'Kinari Bazar*',
 'Kirti Nagar*',
 'Kishan Ganj*',
 'Kotla Mubarak Pur*',
 'Krishna Nagar',
 'Krishna Park*',
 'Kucha Chelan*',
 'Kucha Pandit (Bazar Sita Ram)',
 'Kucha Pati Ram (Pahar Ganj)',
 'Kucha Sohan Lal (Pahar Ganj)',
 'Kuchalal Man (Darya Ganj)',
 'Kunde Walan',
 'Laddu Ghati*',
 'Lajpat Nagar I*',
 'Lajpat Nagar*',
 'Lal Kuan',
 'Loha Mandi Naraina*',
 'Madanpur Khadar*',
 'Madipur*',
 'Mahipal Pur*',
 'Maidan Garhi*',
 'Main Bazar Pahar Ganj*',
 'Malka Ganj*',
 'Malviya Nagar*',
 'Man Singh Road*',
 'Manak Pura',
 'Mandawali*',
 'Mantola*',
 'Masih Garh*',
 'Masjid Moth*',
 'Masood Pur*',
 'Mata Rameshwari Nagar',
 'Mata Sundari Rly Colony (Minto Road)',
 'Mata Sundri Road*',
 'Matia Mahal*',
 'May Fair Garden*',
 'Meena Bazar (Jama Masjid)',
 'Mehrauli*',
 'Militry Road',
 'Minto Road',
 'Mirdard Road*',
 'Mithapur*',
 'Model Basti (Manakpura)',
 'Molarband*',
 'Mori Gate*',
 'Moti Bagh -1*',
 'Moti Nagar*',
 'Motia Bagh*',
 'Motia Khan*',
 'Multani Dhanda',
 'Munirka*',
 'Nabi Karim',
 'Nai Basti*',
 'Nai Sarak*',
 'Nai Wala',
 'Nai Wara*',
 'Naiwala*',
 'Nanak Pura*',
 'Nangal Raya*',
 'Naraina Vihar*',
 'Naraina*',
 'Nawab Ganj*',
 'Naya Bazar*',
 'Nehru Nagar*',
 'Netaji Nagar*',
 'New Friends Colony*',
 'New Moti Nagar*',
 'New Rajinder Nagar',
 'New Ranjit Nagar*',
 'New Rohtak Road*',
 'Niti Bagh*',
 'Nizamuddin West*',
 'North Extn. Area, Pusa Road',
 'Okhla Village*',
 'Old Daryaganj Area( Pataudi House)',
 'Old Rajinder Nagar',
 'Old Subzi Mandi*',
 'Others',
 'Padam Singh Road',
 'Padmini Enclave*',
 'Pahar Ganj',
 'Pahari Bhojla*',
 'Pahari Gajaan',
 'Pahari Imli',
 'Pai Walan (Chandni Chowk)',
 'Pamposh Enclave*',
 'Pandara Road*',
 'Pandav Nagar*',
 'Pant Nagar*',
 'Parsad Nagar*',
 'Partap Nagar*',
 'Paschim Puri*',
 'Paschim Vihar*',
 'Patel Nagar*',
 'Peshwa Road*',
 'Pitampura*',
 'Prashad Nagar',
 'Prem Nagar*',
 'Press Area (Darya Ganj)',
 'Prithvi Raj Road*',
 'Punjabi Bagh Extension*',
 'Punjabi Bagh*',
 'Punjabi Basti*',
 'Pusa Institute*',
 'Pusa Road',
 'Pyare Lal Road',
 'Qamra Bangush',
 'Qasab Pura*',
 'Qutab Road*',
 'Raigar Pura*',
 'Rajender Nagar*',
 'Rajendra Place',
 'Rajinder Nagar',
 'Rajindra Park, Pusa Road',
 'Rajouri Garden*',
 'Rajpur Road*',
 'Rakab Ganj (Darya Ganj)',
 'Rakab Ganj*',
 'Ram Nagar',
 'Rama Krishna Ashram Marg*',
 'Ramesh Nagar*',
 'Ramjas Road*',
 'Rampura*',
 'Rangpuri*',
 'Rani Jhansi Road',
 'Ravi Nagar*',
 'Regar Pura*',
 'Reghar Pura',
 'Rishi Nagar*',
 'Rodgran*',
 'Rohini*',
 'Roshanara Road*',
 'Rouse Avenue (I.T.O.)',
 'Sabzi Mandi*',
 'Sadar Bazar*',
 'Sadar Nala Road*',
 'Sadar Thana Road*',
 'Safdarjung Enclave*',
 'Sainik Farm*',
 'Samalka*',
 'Sanjay Amar Colony (Minto Road)',
 'Sanjay Nagar*',
 'Sant Nagar*',
 'Sarai Kale Khan*',
 'Sardar Patel Marg*',
 'Sarita Vihar*',
 'Sarojini Nagar*',
 'Sarvodaya Enclave*',
 'Sarvpriya Vihar*',
 'Sat Nagar',
 'Savitri Nagar*',
 'Sewa Nagar*',
 'Shadi Khampur*',
 'Shah Ganj (Bazar Sita Ram)',
 'Shahbad Mohammad Pur*',
 'Shahpur Jat*',
 'Shahurpur*',
 'Shakti Nagar*',
 'Shakur Basti*',
 'Shalimar Bagh*',
 'Shastri Park(Beadan Pura)',
 'Sheesh Mahal (Darya Ganj)',
 'Shiv Nagar*',
 'Shivaji Park*',
 'Shivalik*',
 'Shora Kothi (Pahar Ganj)',
 'Shora Kothi*',
 'Shyam Nagar*',
 'Siddi Pura',
 'Sidharth Basti*',
 'Sita Ram Bazar*',
 'South Patel Nagar*',
 'Sriniwaspuri*',
 'Subhash Nagar',
 'Subzi Mandi*',
 'Sui Walan',
 'Suiwalan*',
 'Sukhdev Vihar*',
 'Sunder Nagar*',
 'Sunder Vihar*',
 'Sunlight Colony*',
 'Swami Ramtirth Nagar (manakpura)',
 'Tagore Garden*',
 'Tagore Road*',
 'Taimoor Nagar*',
 'Tajpur*',
 'Thomson Road*',
 'Tibbia College*',
 'Tilak Bridge*',
 'Tilak Nagar*',
 'Timarpur*',
 'Tiraha Behram Khan',
 'Tiraha Behram Khan*',
 'Tis Hazari*',
 'Turkman Gate',
 'Uday Park*',
 'Uttam Nagar*',
 'Varinder Nagar*',
 'Vasant Kunj*',
 'Vijay Nagar*',
 'Vikas Puri*',
 'Vikram Nagar (Minto Road)',
 'Vikram Nagar*',
 'W.E.A. Karol Bagh',
 'Wazir Nagar*',
 'West Patel Nagar*',
 'Yamuna Bazar*',
 'Yusuf Sarai*',
 'Zakhira*',
 'Zakir Nagar*',
 'Zamrud Pur*']
```

Observation: Here we can see the list for locations we have under Central Asaf Ali (SR III).

In [4]:
```python
len(locations) # Lenghth Of location, we have 392 locations present.
```

Out[4]:
```
392
```

In [7]:
```python
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import Select
from time import sleep
import pandas as pd
import numpy as np
from bs4 import BeautifulSoup
import urllib.request
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import NoSuchElementException, ElementClickInterceptedException
from selenium.webdriver.chrome.service import Service


# Initialize the webdriver
driver = webdriver.Chrome(executable_path=r"D:/Malhar H Jadhav/Devlopment/chromedriver.exe")

# Loop through all locations
for i in locations:
    # Open the website and select SRO, locality, and year
    driver.get("http://esearch.delhigovt.nic.in/Complete_search.aspx")
    Select(driver.find_element(By.ID, "ctl00_ContentPlaceHolder1_ddl_sro_s")).select_by_visible_text("Central -Asaf Ali (SR III)")
    sleep(1)
    Select(driver.find_element(By.ID, "ctl00_ContentPlaceHolder1_ddl_loc_s")).select_by_visible_text(i)
    Select(driver.find_element(By.ID, "ctl00_ContentPlaceHolder1_ddl_year_s")).select_by_visible_text("2021-2022")

    # Wait for user to manually enter captcha
    sleep(1)
    # Click the search button
    try:
        driver.find_element(By.ID, "ctl00_ContentPlaceHolder1_btn_search_s").click()

    except ElementClickInterceptedException:
        continue
    # Check if data is available for this location
    try:
        tot_pages = driver.find_element(By.ID, "ctl00_ContentPlaceHolder1_gv_search_ctl13_lblTotalNumberOfPages").get_attribute('innerHTML')
        tot_pages = int(tot_pages)
        cur_page = 1
    except NoSuchElementException:
        # If no data is available for this location, skip it and continue with the next location
        continue

    # Loop through all pages of the search results
    df = pd.DataFrame()
    while cur_page <= tot_pages:
        # Parse the table on the current page and add it to the dataframe
        soup = BeautifulSoup(driver.page_source, 'lxml')
        tables = soup.find_all('table')
        df_temp = pd.read_html(str(tables))[0]
        df_temp.drop(df_temp.tail(1).index, inplace=True)
        df = pd.concat([df, df_temp])

        # Click the "Next" button to go to the next page
        cur_page += 1
        if cur_page <= tot_pages:
            sleep(1)
            element = driver.find_element(By.ID, "ctl00_ContentPlaceHolder1_gv_search_ctl13_Button2")
            driver.execute_script("arguments[0].click();", element)
            sleep(2)




    # Save the data for this location to a CSV file
    df.to_csv(f"{i}.csv", index=False)



# Close the webdriver
driver.quit()
```

```
C:\Users\harsh\AppData\Local\Temp\ipykernel_2792\1367734829.py:17: DeprecationWarning: executable_path has been deprecated, please pass in a Service object
  driver = webdriver.Chrome(executable_path=r"D:/Malhar H Jadhav/Devlopment/chromedriver.exe")
```

Observation: So, here this code will iterate to all the location and if there's any data present it will stored it as a "Location.csv" and will go to next location. And, if there's no data present in next location it will just ignore and try for another location.

In [1]:
```python
import pandas as pd
import os

# set the folder path where the CSV files are located
folder_path = r"D:\Malhar H Jadhav\Tasks\Propreturns\Location Wise Data"

# create an empty list to store all the dataframes
dfs = []

# loop through all the files in the folder
for file in os.listdir(folder_path):
    # check if the file is a CSV file
    if file.endswith(".csv"):
        # extract the location name from the filename
        location = file[:-4] # remove the .csv extension
        # read the CSV file into a dataframe
        df = pd.read_csv(os.path.join(folder_path, file))
        # add a new column with the location name
        df["Location"] = location
        # append the dataframe to the list of dataframes
        dfs.append(df)

# concatenate all the dataframes into one
result = pd.concat(dfs)
```

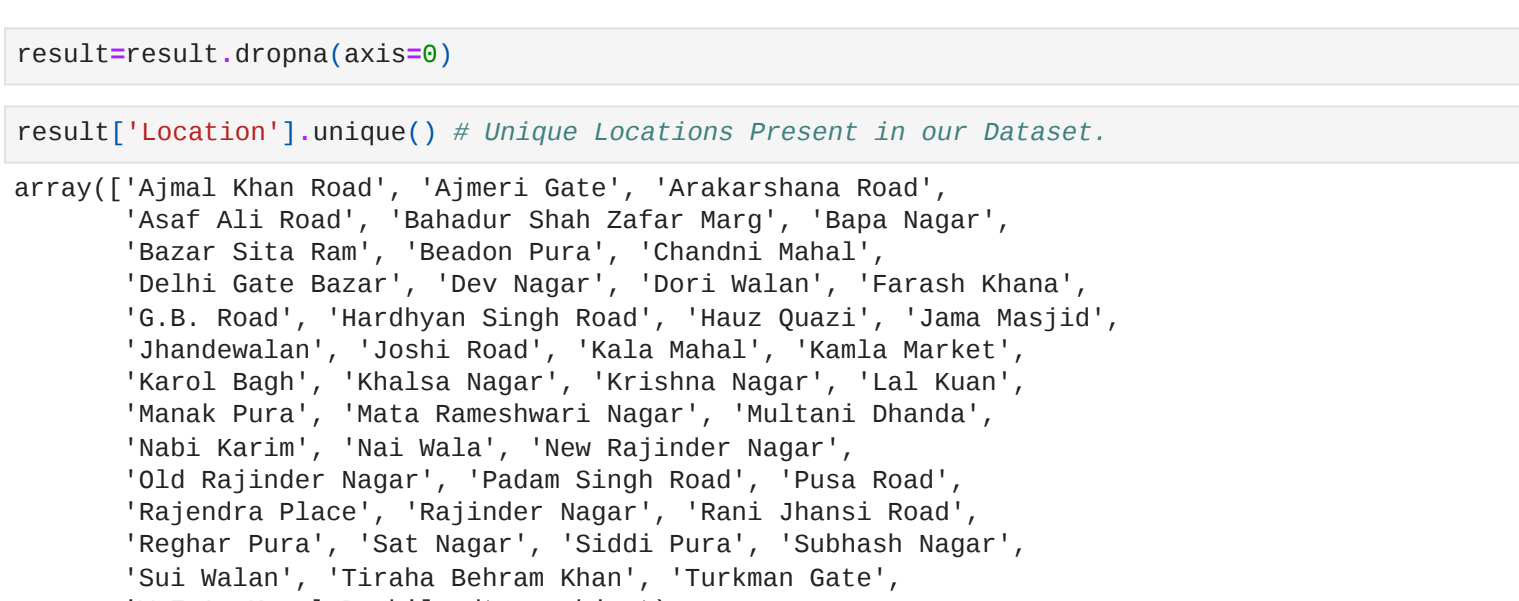
Observation: Here, I'm combining all the DataFrames to one DataFrame.

In [3]:
```python
result.isna().sum() / len(result)*100 # Checking for null
```

Out[3]:
```
Reg.No              52.422089
Reg.Date            52.422089
First Party         52.422089
Second Party        52.422089
Property Address    52.422089
Area                52.422089
Deed Type           52.422089
Property Type       52.422089
Location             0.000000
dtype: float64
```

In [9]:
```python
from IPython.display import Image
Image(filename='A.png')
```

Out[9]:



258	########	VIKAS JAIN	SUSHIL KU	House No.	5.34 Sq. M	SALE,SALE	Commercial
262	########	MOHD YU	ARSHI REH	House No.	32.61 Sq. I	SALE,SALE	Residential
268	########	MOHD YU	MOHD JUI	House No.	32.61 Sq. I	SALE,SALE	Residential
298	14-01-202	RAM KISH/	RAJENDRA	House No.	58.53 Sq. I	RELEASE,R	Commercial
299	14-01-202	RAJENDRA	DEVA NAN	House No.	83.61 Sq. I	RELEASE,R	Commercial
300	14-01-202	ASHOK GA	RAJINDRA	House No.	66.88 Sq. I	RELEASE,R	Commercial

Removing All the null values (Basically, they are not null values they are just extra rows Check above example we have a empty row after every value so removing it)

In [10]:
```python
result=result.dropna(axis=0)
```

In [11]:
```python
result['Location'].unique() # Unique Locations Present in our Dataset.
```

Out[11]:
```
array(['Ajmal Khan Road', 'Ajmeri Gate', 'Arakarshana Road',
       'Asaf Ali Road', 'Bahadur Shah Zafar Marg', 'Bapa Nagar',
       'Bazar Sita Ram', 'Beadon Pura', 'Chandni Mahal',
       'Delhi Gate Bazar', 'Dev Nagar', 'Dori Walan', 'Farash Khana',
       'G.B. Road', 'Hardhyan Singh Road', 'Hauz Quazi', 'Jama Masjid',
       'Jhandewalan', 'Joshi Road', 'Kala Mahal', 'Kamla Market',
       'Karol Bagh', 'Khalsa Nagar', 'Krishna Nagar', 'Lal Kuan',
       'Manak Pura', 'Mata Rameshwari Nagar', 'Multani Dhanda',
       'Nabi Karim', 'Nai Wala', 'New Rajinder Nagar',
       'Old Rajinder Nagar', 'Padam Singh Road', 'Pusa Road',
       'Rajendra Place', 'Rajinder Nagar', 'Rani Jhansi Road',
       'Reghar Pura', 'Sat Nagar', 'Siddi Pura', 'Subhash Nagar',
       'Sui Walan', 'Tiraha Behram Khan', 'Turkman Gate',
       'W.E.A. Karol Bagh'], dtype=object)
```

In [12]:
```python
result['Location'].nunique()
```

Out[12]:
```
45
```

In [13]:
```python
result
```

Out[13]:

	Reg.No	Reg.Date	First Party	Second Party	Property Address	Area	Deed Type	Property Type	Location
0	1096.0	11-02-2021	ADLAKHA ENTERPRISES TH SOHAN LAL ADLAKHA	VEDANT FASHIONS PVT LTD TH ANIL KUMAR SHARMA	House No. 6/17-18,Ajmal Khan Road	18725 Sq. Feet	LEASE,LEASE WITH SECURITY UPTO 10 YEARS	Commercial	Ajmal Khan Road
2	1097.0	11-02-2021	T K ARORA AND SONS AND OTHER TH VIJAY KUMAR ARORA	VEDANT FASHIONS PVT LTD TH ANIL KUMAR SHARMA	House No. 6/18,Ajmal Khan Road	475 Sq. Feet	LEASE,LEASE WITH SECURITY UPTO 10 YEARS	Commercial	Ajmal Khan Road
4	1187.0	12-02-2021	KAPIL KUMAR	AJITESH DUBEY	House No. 5/53,Ajmal Khan Road	495 Sq. Feet	LEASE,LEASE UPTO 5 YEARS	Commercial	Ajmal Khan Road
6	1359.0	16-02-2021	SHOBHA RANI MEHRA	SHIVALIK INTERNATIONAL TH ASHUTOSH BHUWALKA	House No. 13/29-30,Ajmal Khan Road	450 Sq. Feet	LEASE,LEASE WITH SECURITY UPTO 5 YEARS	Commercial	Ajmal Khan Road
8	1605.0	23-02-2021	SINDHU BHALLA AND MUKESH KUMAR BHALLA	IIFL HOME FINANCE LIMITED TH ARJUN	House No. 16-A/13 PLOT NO 13,Ajmal Khan Road	1950 Sq. Feet	MORTGAGE,MORTGAGE WITHOUT POSSESSION	Residential	Ajmal Khan Road
...	...	...	...	...	...	...	...	...	...
2484	11563.0	31-12-2021	BHAGAT JI GARMENTS PVT LTD TH RITU GUPTA	MANMOON SINGH KALRA AND OTHERS	House No. 10546,W.E.A. Karol Bagh	215.7 Sq. Meter	SALE,SALE WITHIN MC AREA	Commercial	W.E.A. Karol Bagh
2486	11640.0	31-12-2021	MUKESH KUMAR GUPTA	AJIT SINGH	House No. 13/2,W.E.A. Karol Bagh	66.92 Sq. Meter	SALE,SALE WITHIN MC AREA	Commercial	W.E.A. Karol Bagh
2488	11719.0	31-12-2021	KANWALJIT KAUR	BIBHA DEVI	House No. 10178,W.E.A. Karol Bagh	72 Sq. Meter	SALE,SALE WITHIN MC AREA	Residential	W.E.A. Karol Bagh
2490	11725.0	31-12-2021	RAJESH WADHWA	CHHAYA WADHWA	House No. C-3/43,W.E.A. Karol Bagh	131.26 Sq. Meter	GIFT,GIFT WITH IN MC AREA	Residential	W.E.A. Karol Bagh
2492	11726.0	31-12-2021	MADHU SUDAN SAINI	LALLAN KUMAR	House No. 10178,W.E.A. Karol Bagh	22.3 Sq. Meter	SALE,SALE WITHIN MC AREA	Residential	W.E.A. Karol Bagh

10244 rows × 9 columns

In [8]:
```python
# save the combined dataframe to a new CSV file
result.to_csv("combined_data.csv", index=False)
```

Observation: Saving the DataFrame into a CSV File for futher use.

Thank You For This Oppurtunity

In [ ]: