Data Crawling Assesment from TM&RD.

Web Crawling is a term used to describe the use of a program or algorithm to extract and process
Whether you are a data scientist, engineer, or anybody who analyzes large amounts of datasets

Import the required package

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import datetime
import time
```

```
titleList = []
priceList = []
areaList = []
typesList = []
sizeList = []
bedroomList = []
datesList =[]
timesStampList = []
```

For this assesment I'm crawling only the few pages of the web pages since.

To scrape more we could increase the while loop to scrape more than 100 pages etc.

```
i = 1

url = "https://www.mudah.my/Selangor/Apartments-for-sale-2020?lst=" +str(i)+ "&fs=1
```

```
while i<=15:

    page = requests.get(url)
    soup = BeautifulSoup(page.content, 'lxml')

    title = soup.find_all('h2', {'class':'list_title'})
    price = soup.find_all('div', {'class':'ads_price'})
    area = soup.find_all('div', {'class':'area'})
    types = soup.find_all('div', {'class':'apartments'})
    size = soup.find_all('div', {'class':'apartmentAndLandAndRoomsAndNew-Properties
    bedroom = soup.find_all('div', {'class':'bedroom'})
    dates = soup.find_all('div', {'class':'location bottom_info'})
    timesstamp = soup.findAll('div', {'class':'_timestamp'})

    timesstamp = [k.text for k in timesstamp]
    timesstamp = timesstamp[:-1]
```

```
    title = [k.text for k in title]
    title = title[:-1]

    price = [k.text for k in price]
    price = price[:-1]

    area = [k.text for k in area]
    area = area[:-1]

    types = [k.text for k in types]
    types = types[:-1]

    size = [k.text for k in size]
    size = size[:-1]

    bedroom = [k.text for k in bedroom]
    bedroom = bedroom[:-1]

    dates = [k.text for k in dates]
    dates = dates[:-1]

    titleList += title
    priceList += price
    areaList += area
    typesList += types
    sizeList += size
    bedroomList += bedroom
    datesList += dates

    print("number of pages that been scraped : ", i)
    i = i+1
```

```
number of pages that been scraped :  1
number of pages that been scraped :  2
number of pages that been scraped :  3
number of pages that been scraped :  4
number of pages that been scraped :  5
number of pages that been scraped :  6
number of pages that been scraped :  7
number of pages that been scraped :  8
number of pages that been scraped :  9
number of pages that been scraped :  10
number of pages that been scraped :  11
number of pages that been scraped :  12
number of pages that been scraped :  13
number of pages that been scraped :  14
number of pages that been scraped :  15
```

```
df_title = pd.DataFrame()
df_price = pd.DataFrame()
df_area = pd.DataFrame()
df_types = pd.DataFrame()
df_size = pd.DataFrame()
df_bedroom = pd.DataFrame()
df_dates = pd.DataFrame()
```

```
print(len(titleList))
print(len(priceList))
print(len(areaList))
print(len(typesList))
print(len(sizeList))
print(len(bedroom))
print(len(datesList))


df_dates['Date'] = datesList
df_area['Area'] = areaList
df_title['Title'] = titleList
df_types['Type'] = typesList
df_bedroom['Bedroom'] = bedroomList
df_size['Size'] = sizeList
df_price['Price'] = priceList
```

```
600
600
585
630
630
42
600
```

```
df_apartements = pd.concat([df_dates, df_area, df_title,df_types, df_bedroom, df_si
df_apartements.head(10)
```

	Date	Area	Title	Type	Bedroom	Size	Price
0	Today, 01:55 Sepang	Sepang	Flat dahlia, taman dahlia, bandar baru salak...	Apartments	3 Bedrooms	651 sq.ft	RM 120 000
1	Today, 01:45 Sepang	Sepang	Sepang Best Selling Condo Near KLIA! Call Me!	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
2	Today, 01:45 Petaling Jaya	Petaling Jaya	Paradesa Rustica Condominium	Apartments	3 Bedrooms	1045 sq.ft	RM 509 999
3	Today, 01:40 Cyberjaya	Cyberjaya	Best Investment at Cybersouth! Hot Selling !...	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
4	Today, 01:05 Petaling Jaya	Petaling Jaya	100% Loan , Free MOT , 1km to Jaya33 , Unive...	Apartments	2 Bedrooms	750 sq.ft	RM 498 000
5	Today, 01:05 Petaling Jaya	Petaling Jaya	First Home Buyer , Free MOT , 100% Loan + Fr...	Apartments	3 Bedrooms	850 sq.ft	RM 570 000
6	Today, 00:41 Bandar Sunway	Bandar Sunway	Bandar sunway, sri subang apartment, tingkat...	Apartments	3 Bedrooms	785 sq.ft	RM 220 000
	Today, 00:41		KENANGA APARTMENT TAMAN		3	800	RM

```
df_apartements.isnull().sum()
```

```
Date       30
Area       45
Title      30
Type        0
Bedroom     0
Size        0
Price      30
dtype: int64
```

```
df_apartements.dropna(inplace= True)
df_apartements.shape
```

```
(585, 7)
```

```
df_apartements
```

	Date	Area	Title	Type	Bedroom	Size	Price
0	Today, 01:55 Sepang	Sepang	Flat dahlia, taman dahlia, bandar baru salak...	Apartments	3 Bedrooms	651 sq.ft	RM 120 000
1	Today, 01:45 Sepang	Sepang	Sepang Best Selling Condo Near KLIA! Call Me!	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
2	Today, 01:45 Petaling Jaya	Petaling Jaya	Paradesa Rustica Condominium	Apartments	3 Bedrooms	1045 sq.ft	RM 509 999
3	Today, 01:40 Cyberjaya	Cyberjaya	Best Investment at Cybersouth! Hot Selling !...	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
4	Today, 01:05 Petaling Jaya	Petaling Jaya	100% Loan , Free MOT , 1km to Jaya33 , Unive...	Apartments	2 Bedrooms	750 sq.ft	RM 498 000
...	...	...	...	...	...	...	...
580	Yesterday, 21:45 Shah Alam	Shah Alam	Apartment Desa Subang, Jalan Waruna, Seksyen...	Apartments	3 Bedrooms	603 sq.ft	RM 190 000
	Yesterday, 21:45		Apartment Service		2	516	RM

```
#df_apartements.to_csv('datasets/exportDataFrames.csv', index=False, header=True)
```

```
df_apartements.isnull().sum()
```

```
Date       0
Area       0
Title      0
Type       0
Bedroom    0
Size       0
Price      0
dtype: int64
```

```
df_apartements.isna().sum()
```

```
Date       0
Area       0
Title      0
Type       0
Bedroom    0
Size       0
Price      0
dtype: int64
```

```
df_apartements.isna().any()
```

```
Date       False
Area       False
Title      False
Type       False
Bedroom    False
Size       False
Price      False
dtype: bool
```

```
df_apartements.isna().any(axis = None)
```

```
False
```

```
df_apartements['Date'][1:]
```

```
1                    Today, 01:45   Sepang
2              Today, 01:45   Petaling Jaya
3                  Today, 01:40   Cyberjaya
4              Today, 01:05   Petaling Jaya
5              Today, 01:05   Petaling Jaya
                          ...
580            Yesterday, 21:45   Shah Alam
581            Yesterday, 21:45   Shah Alam
582             Yesterday, 21:42   Semenyih
583               Yesterday, 21:40   Kajang
584      Yesterday, 21:40   Seri Kembangan
Name: Date, Length: 584, dtype: object
```

```
df_apartements["Date"] = df_apartements["Date"].str.split(n = 1).str[0]
#df_apartements["dates"] = df_apartements["Date"].str.split(n = 1).str[1]
df_apartements
```

	Date	Area	Title	Type	Bedroom	Size	Price
0	Today,	Sepang	Flat dahlia, taman dahlia, bandar baru salak...	Apartments	3 Bedrooms	651 sq.ft	RM 120 000
1	Today,	Sepang	Sepang Best Selling Condo Near KLIA! Call Me!	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
2	Today,	Petaling Jaya	Paradesa Rustica Condominium	Apartments	3 Bedrooms	1045 sq.ft	RM 509 999
3	Today,	Cyberjaya	Best Investment at Cybersouth! Hot Selling !...	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
4	Today,	Petaling Jaya	100% Loan , Free MOT , 1km to Jaya33 , Unive...	Apartments	2 Bedrooms	750 sq.ft	RM 498 000
...	...	...	...	...	...	...	...
			Apartment Desa		3	603	RM

```
df_apartements['Date'] = df_apartements['Date'].str.replace(',', '')
```

```
df_apartements
```

	Date	Area	Title	Type	Bedroom	Size	Price
0	Today	Sepang	Flat dahlia, taman dahlia, bandar baru salak...	Apartments	3 Bedrooms	651 sq.ft	RM 120 000
1	Today	Sepang	Sepang Best Selling Condo Near KLIA! Call Me!	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
2	Today	Petaling Jaya	Paradesa Rustica Condominium	Apartments	3 Bedrooms	1045 sq.ft	RM 509 999
3	Today	Cyberjaya	Best Investment at Cybersouth! Hot Selling !...	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
4	Today	Petaling Jaya	100% Loan , Free MOT , 1km to Jaya33 , Unive...	Apartments	2 Bedrooms	750 sq.ft	RM 498 000
...	...	...	...	...	...	...	...
			Apartment Desa		3	603	RM

```
#del df_apartements['date']
```

```
from google.colab import drive
```

```
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
```

```
!ls
```

```
drive  sample_data
```

```
cd /content/drive/My Drive/web-scrapping
```

```
/content/drive/My Drive/web-scrapping
```

```
!ls
```

```
exportDataFrames.csv  exportDataFrames.gsheet  exportDatasets.csv
```

```
df_apartements.to_csv('/content/drive/My Drive/web-scrapping/exportDataset.csv', in
```

```
df_apartements.dropna(inplace= True)
df_apartements.shape
```

```
(585, 7)
```

```
df_apartements
```

	Date	Area	Title	Type	Bedroom	Size	Price
0	Today,	Sepang	Flat dahlia, taman dahlia, bandar baru salak...	Apartments	3 Bedrooms	651 sq.ft	RM 120 000
1	Today,	Sepang	Sepang Best Selling Condo Near KLIA! Call Me!	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
2	Today,	Petaling Jaya	Paradesa Rustica Condominium	Apartments	3 Bedrooms	1045 sq.ft	RM 509 999
3	Today,	Cyberjaya	Best Investment at Cybersouth! Hot Selling !...	Apartments	3 Bedrooms	660 sq.ft	RM 238 000
4	Today,	Petaling Jaya	100% Loan , Free MOT , 1km to Jaya33 , Unive...	Apartments	2 Bedrooms	750 sq.ft	RM 498 000
...	...	...	...	...	...	...	...
			Apartment Desa		3	603	RM

```
df_apartements.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 585 entries, 0 to 584
Data columns (total 7 columns):
Date       585 non-null object
Area       585 non-null object
Title      585 non-null object
Type       585 non-null object
Bedroom    585 non-null object
Size       585 non-null object
Price      585 non-null object
dtypes: object(7)
memory usage: 36.6+ KB
```

```
from datetime import datetime
```

```
# df_apartements['Date'] =  pd.to_datetime(df_apartements['Date'],
#                                format='')
```

```
import matplotlib.pyplot as plt
%matplotlib inline
```

```
cleaned_df = df_apartements[~df_apartements['Title'].str.contains(' naza ', case=Fa
cleaned_df.shape
```

```
(585, 7)
```

```
cleaned_df['Price'].replace(regex=True,inplace=True,to_replace=r'\D',value=r'')
cleaned_df['Price'] = pd.to_numeric(cleaned_df['Price'])
```

```
cleaned_df
```

```
location_group = cleaned_df.groupby(['Area'])['Title'].count().reset_index()

location_average = cleaned_df.groupby(['Area'])['Price'].mean().reset_index()


display(location_group, round(location_average, 2))
```

	Area	Title
0	Ampang	15
1	Bandar Sunway	15
2	Bangi	15
3	Bukit Jelutong	15
4	Cyberjaya	30
5	Kajang	75
6	Kota Damansara	15
7	Petaling Jaya	60
8	Puchong	15
9	Rawang	30
10	Semenyih	30
11	Sepang	30
12	Seri Kembangan	60
13	Shah Alam	150
14	Subang Bestari	15
15	USJ	15

	Area	Price
0	Ampang	216846.67
1	Bandar Sunway	296599.93
2	Bangi	314599.93
3	Bukit Jelutong	234713.33
4	Cyberjaya	250223.30
5	Kajang	316329.29
6	Kota Damansara	352666.60
7	Petaling Jaya	262639.95
8	Puchong	301866.60
9	Rawang	281533.33
10	Semenyih	311766.67
11	Sepang	241513.33
12	Seri Kembangan	227668.33
13	Shah Alam	272993.99
14	Subang Bestari	367733.27
15	USJ	349933.27

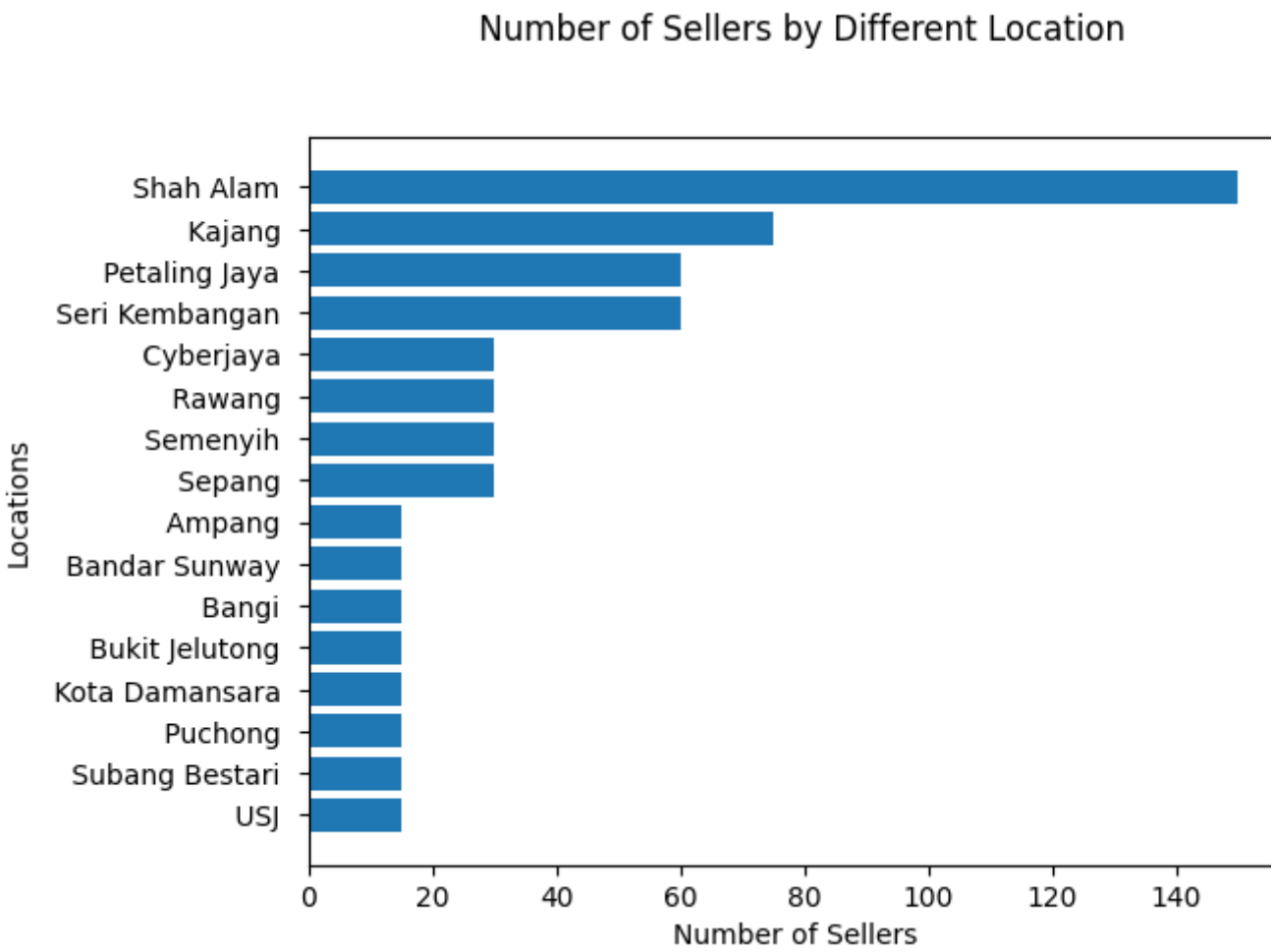
```
location_group = location_group.sort_values('Title', ascending=False).reset_index(d
location_group.Title

plt.rcdefaults()
fix,ax = plt.subplots()

locations = location_group.Area.tolist()
y_pos = np.arange(len(locations))
names = location_group.Title.tolist()

ax.barh(y_pos,names)
ax.set_yticks(y_pos)
ax.set_yticklabels(locations)
ax.invert_yaxis()
ax.set_xlabel('Number of Sellers')
ax.set_ylabel('Locations')
ax.set_title('\n\n Number of Sellers by Different Location\n\n')

plt.show()
```



To be continued