In [ ]:
```
import requests
import urllib.request
from bs4 import BeautifulSoup
import pandas as pd
```

In [ ]:
```
class Cars:
    def __init__(self, mark_, price_, produced_year_, imported_year_, distance_, motor_volume_, color_, type_, hurd,
        self.mark = mark_
        self.price = price_
        self.produced_year = produced_year_
        self.imported_year = imported_year_
        self.distance = distance_
        self.motor_volume = motor_volume_
        self.color = color_
        self.type = type_
        self.hurd = hurd
        self.hodolguur = hodolguur_
```

In [ ]:
```
baseurl = "https://www.unegui.mn/avto-mashin/-avtomashin-zarna/?page="
car_list = []
for i in range(1, 200):
    url = baseurl + str(i)

    response = requests.get(url)
    if response.status_code != 200:
        print(response.status_code)
        print('error', url)
        continue
    soup = BeautifulSoup(response.text, "html.parser")

    li_list = soup.find_all("div", {"class": "swiper-wrapper"})
    for li in li_list:
        a = li.find('a')
        car_url = "https://www.unegui.mn" + a['href']
        #print(car_url)
        car_list.append(car_url)
```

In [ ]:
```
def findFeature(li_list, header):
    ref = None
    for li in li_list:
        text = li.text.strip()
        if text.startswith(header):
            ref = text[len(header):].strip()
            break
    return ref
```

In [ ]:
```
print(len(car_list))
car_set = set(car_list)
print(len(car_set))
```

```
11940
11940
```

In [ ]:
```
it = 0
cars_data = []
for url in car_set:
    it += 1
    response = requests.get(url)
    if response.status_code != 200:
        print(response.status_code)
        print('error', url)
        continue
    soup = BeautifulSoup(response.text, "html.parser")

    mark = soup.find("h1", {"class": "title-announcement"}).text.strip()
    mark = str(mark.split(',')[0])
    price = soup.find("div", {"class": "announcement-price__cost"}).text.strip()
    price = float(price.split('сая')[0])
    li_class = soup.find_all("li")
    produced_year = findFeature(li_class, "Үйлдвэрлэсэн он:")
    imported_year = findFeature(li_class, "Орж ирсэн он:")
    distance = findFeature(li_class, "Явсан:")
    motor_volume = findFeature(li_class, "Мотор багтаамж:")
    color = findFeature(li_class, "Дотор өнгө:")
    type = findFeature(li_class, "Төрөл:")
    hurd = findFeature(li_class, "Хүрд:")
    hodolguur = findFeature(li_class, "Хөдөлгүүр:")
```

```
    #print(title, price, p_year, i_year, distance, motor, color)
    car = Cars(mark, price, produced_year, imported_year, distance, motor_volume, color, type, hurd, hodolguur)
    cars_data.append(car.__dict__)
    #print(car)
```

In [ ]:
```
df = pd.DataFrame(cars_data)
df.to_csv('car_info.tsv', sep="\t", index=False)
```

unegui.mn - ээс 11940-н авто машинтай холбоотой зар байсан. Гэхдээ scrap хийхэд нэлээн удаж хийгдсэн, тэрнээс болоод зөвхөн 5076 машины зар авч чадлаа. Файл нь доорх холбоост байгаа.

https://drive.google.com/file/d/1CKWn_97UA7gUfx1THx50cbMwWRAbSAvS/view?usp=sharing

Nomin.mn сайтаас электроник барааны зарын мэдээллийг татсан. Гэвч энэ 6-н баганатай.

In [ ]:
```
class Product:
    def __init__(self, name_, brand_, sale_day, sale_price_, sale_percent_, additionally_):
        self.name = name_
        self.brand = brand_
        self.sale_day = sale_day
        self.sale_price = sale_price_
        self.sale_percent = sale_percent_
        self.additionally = additionally_
```

In [ ]:
```
baseurl = 'https://enomin.mn/t/6011?page='
product_list = []
for i in range(1, 63):
    url = baseurl + str(i)
    #print(url)
    response = requests.get(url)
    if response.status_code != 200:
        print(response.status_code)
        print('error ', url)
        continue
    #print(response.text)
    soup = BeautifulSoup(response.text, "html.parser")
    li_list = soup.find_all("div", {"class": "MuiBox-root css-1efcy4n"})
    for li in li_list:
        a = li.find('a')
```

```
        product_url = "https://enomin.mn" + a['href']
        #print(product_url)
        product_list.append(product_url)
```

In [ ]:
```
len(product_list)
```

Out[ ]: 1860

In [ ]:
```
def first2(s):
    return s[:2]
```

In [ ]:
```
product_data = []
it  = 0
for url in product_list:
    it += 1
    #print(url)
    response = requests.get(url)
    if response.status_code != 200:
        #print(response.status_code)
        print('error ', url)
        continue
    soup = BeautifulSoup(response.text, "html.parser")
    name = soup.find("h1", {"class": "MuiBox-root css-1ozh2ah"}).text.strip()
    brand = soup.find("h6", {"class": "MuiBox-root css-jyp6ua"}).text.strip()
    sale_day = soup.find("div", {"_2EBbg"}).text.strip()
    sale_day = first2(sale_day)
    sale_price = soup.find("h2", {"class": "MuiBox-root css-x5yzb2"}).text.strip()
    sale_percent = soup.find("span", {"class": "MuiChip-label MuiChip-labelSmall css-19imqg1"}).text.strip()
    additionally = soup.find("div", {"class": "MuiBox-root css-10ibhvy"}).text.strip()

    products = Product(name, brand, sale_day, sale_price, sale_percent, additionally)
    product_data.append(products.__dict__)
    #print(name, brand, sale_day, sale_price, description, weight)
```

In [ ]:
```
df = pd.DataFrame(product_data)
df.to_csv('electronic.tsv', sep="\t", index=False)
```

Мөн 1860 зараас 823-ыг нь татаж чадсан. Энэ удаад татагдаж байсан ч дундаас нь алдаа заагаад хагас нь татагдаагүй. Тэр ямар учиртай юм бол багшаа.

https://drive.google.com/file/d/1E13u17TFsmz8LlrmPIPq0vGjAjdeeT_B/view?usp=sharing

Мөн багшаа зарим нэгэн сайтууд руу ороход тэдний вэб inspect хийгдэхгүй байсан.

1. Тэдний HTML кодыг харах ямар нэгэн арга байгаа юу?
2. Тэр HTML код ямар учраас inspect хийгдэхгүй байгаа вэ? Ямар арга ашиглаж тийм болгодог бол? Жишээ нь : emonos.mn сайт