Chapter 1 - Ex1: Tiền xử lý dữ liệu - Điểm thi THPT Quốc Gia 2016

- Cho tập tin Diemthi_thpt_quocgia_2016.xlsx chứa bộ dữ liệu điểm thi THPT Quốc Gia năm 2016 của gần 35.000 thí sinh. ### Yêu cầu:
- Đọc dữ liệu, xem thông tin dữ liệu

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34826 entries, 0 to 34825
Data columns (total 6 columns):
SOBAODANH     34826 non-null object
HO_TEN        34826 non-null object
NGAY_SINH     34826 non-null object
TEN_CUMTHI    34826 non-null object
GIOI_TINH     34826 non-null object
DIEM_THI      34826 non-null object
```

- Biết danh sách các môn thi là: "Toán","Ngữ văn","Địa lí","Tiếng Anh","Sinh học","Vật lí","Hóa học","Lịch sử". Một thí sinh chỉ thi các môn bắt buộc chung còn các môn tự chọn có thể khác nhau.
- Với dữ liệu hiện tại, cột DIEM_THI là chuỗi chứa điểm thi của tất cả các môn mà một thí sinh thi.

	SOBAODANH	HO_TEN	NGAY_SINH	TEN_CUMTHI	GIOI_TINH	DIEM_THI
0	018000001	DƯƠNG VIỆT AN	12/03/1998	Sở GDĐT Bắc Giang	Nam	Toán: 2.00 Ngữ văn: 5.50 Lịch sử: 3....
1	018000002	ĐỖ VĂN AN	09/12/1998	Sở GDĐT Bắc Giang	Nam	Toán: 5.50 Ngữ văn: 5.25 Địa lí: 5.5...
2	018000003	ĐỖ XUÂN AN	12/08/1997	Sở GDĐT Bắc Giang	Nam	Toán: 4.50 Ngữ văn: 5.50 Địa lí: 3.7...
3	018000004	ĐẶNG PHÚC AN	19/03/1998	Sở GDĐT Bắc Giang	Nữ	Toán: 3.00 Ngữ văn: 6.00 Địa lí: 5.5...
4	018000005	ĐẶNG VĂN AN	25/10/1998	Sở GDĐT Bắc Giang	Nam	Toán: 2.25 Ngữ văn: 4.75 Địa lí: 5.2...

- Và như vậy thì chúng ta sẽ không phân tích được điểm thi của thí sinh. Do đó, việc đầu tiên là phải tiền xử lý dữ liệu. Từ dữ liệu trong cột DIEM_THI, hãy tạo ra các cột tương ứng với danh sách các môn thi nói trên và đưa điểm của thí sinh từ chuỗi vào các cột, môn nào thí sinh không thi thì sẽ để NaN.

	SOBAODANH	HO_TEN	NGAY_SINH	TEN_CUMTHI	GIOI_TINH	DIEM_THI	Toán	Ngữ văn	Địa lí	Tiếng Anh	Sinh học	Vật lí	Hóa học	Lịch sử
0	018000001	DƯƠNG VIỆT AN	12/03/1998	Sở GDĐT Bắc Giang	Nam	Toán: 2.00 Ngữ văn: 5.50 Lịch sử: 3....	2.00	5.50	5.00	NaN	NaN	NaN	NaN	3.0
1	018000002	ĐỖ VĂN AN	09/12/1998	Sở GDĐT Bắc Giang	Nam	Toán: 5.50 Ngữ văn: 5.25 Địa lí: 5.5...	5.50	5.25	5.50	3.68	NaN	NaN	NaN	NaN
2	018000003	ĐỖ XUÂN AN	12/08/1997	Sở GDĐT Bắc Giang	Nam	Toán: 4.50 Ngữ văn: 5.50 Địa lí: 3.7...	4.50	5.50	3.75	2.25	NaN	NaN	NaN	NaN
3	018000004	ĐẶNG PHÚC AN	19/03/1998	Sở GDĐT Bắc Giang	Nữ	Toán: 3.00 Ngữ văn: 6.00 Địa lí: 5.5...	3.00	6.00	5.50	1.50	NaN	NaN	NaN	NaN
4	018000005	ĐẶNG VĂN AN	25/10/1998	Sở GDĐT Bắc Giang	Nam	Toán: 2.25 Ngữ văn: 4.75 Địa lí: 5.2...	2.25	4.75	5.25	2.00	NaN	NaN	NaN	NaN

- Hãy vẽ biểu đồ phân phối tần suất điểm thi, mỗi điểm thi là một biểu đồ, nhận xét trên từng biểu đồ: các thống kê mô tả, phân phối chuẩn hay nghiêng? Đường cong cao hơn hay thấp hơn phân phối chuẩn…
- Lưu dữ liệu điểm thi sau khi đã chuẩn hóa để sử dụng.

In [ ]:

```
# from google.colab import drive
# drive.mount("/content/gdrive", force_remount=True)
# %cd '/content/gdrive/My Drive/LDS5/Practice/Chapter1/'
```

```
Mounted at /content/gdrive
```

In [1]:

```
import numpy as np
import pandas as pd
import re
```

In [2]:

```
df = pd.read_excel("Diemthi_thpt_quocgia_2016.xlsx",
                   sheet_name="Export Worksheet")
```

In [3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34826 entries, 0 to 34825
Data columns (total 6 columns):
SOBAODANH     34826 non-null object
HO_TEN        34826 non-null object
NGAY_SINH     34826 non-null object
TEN_CUMTHI    34826 non-null object
GIOI_TINH     34826 non-null object
DIEM_THI      34826 non-null object
dtypes: object(6)
memory usage: 1.6+ MB
```

In [5]:

```
df.head()
```

Out[5]:

	SOBAODANH	HO_TEN	NGAY_SINH	TEN_CUMTHI	GIOI_TINH	DIEM_THI
0	018000001	DƯƠNG VIỆT AN	12/03/1998	Sở GDĐT Bắc Giang	Nam	Toán: 2.00 Ngữ văn: 5.50 Lịch sử: 3....
1	018000002	ĐỖ VĂN AN	09/12/1998	Sở GDĐT Bắc Giang	Nam	Toán: 5.50 Ngữ văn: 5.25 Địa lí: 5.5...
2	018000003	ĐỖ XUÂN AN	12/08/1997	Sở GDĐT Bắc Giang	Nam	Toán: 4.50 Ngữ văn: 5.50 Địa lí: 3.7...
3	018000004	ĐẶNG PHÚC AN	19/03/1998	Sở GDĐT Bắc Giang	Nữ	Toán: 3.00 Ngữ văn: 6.00 Địa lí: 5.5...
4	018000005	ĐẶNG VĂN AN	25/10/1998	Sở GDĐT Bắc Giang	Nam	Toán: 2.25 Ngữ văn: 4.75 Địa lí: 5.2...

In [ ]:

```
ds_mon_thi=("Toán","Ngữ văn","Địa lí","Tiếng Anh",
            "Sinh học","Vật lí","Hóa học","Lịch sử")
```

In [ ]:

```
df_target = pd.DataFrame(columns=list(ds_mon_thi))
```

In [ ]:

```
def tach_diem(str_diem):
    dict_diem={}
    for mon_thi in ds_mon_thi:
        #str_diem="Toán:   6.00   Ngữ văn:   7.00   Hóa học:   5.40   Sinh học:   6.00
Tiếng Anh:   2.50"
        chuoi_tim = mon_thi+':\s\s\s\d.\d\d'
        #print(chuoi_tim)
        match = re.search(chuoi_tim, str_diem)
        # If-statement after search() tests if it succeeded
        if match:
            tmp_list = list(match.group().partition(":   "))
            # print(tmp_list)
            dict_diem[mon_thi] = [float(tmp_list[2])]
        else:
            dict_diem[mon_thi] = [np.nan]
    return pd.DataFrame(dict_diem)
```

In [ ]:

```
for item in range(df.shape[0]):
    str_diem = df["DIEM_THI"][item]
    temp = tach_diem(str_diem)
    df_target = pd.concat([df_target, temp], sort=False,
                          ignore_index=True)
```

In [ ]:

```
#df_target = df_target.reset_index()
```

In [ ]:

```
df_target.head()
```

Out[ ]:

	Toán	Ngữ văn	Địa lí	Tiếng Anh	Sinh học	Vật lí	Hóa học	Lịch sử
0	2.00	5.50	5.00	NaN	NaN	NaN	NaN	3.0
1	5.50	5.25	5.50	3.68	NaN	NaN	NaN	NaN
2	4.50	5.50	3.75	2.25	NaN	NaN	NaN	NaN
3	3.00	6.00	5.50	1.50	NaN	NaN	NaN	NaN
4	2.25	4.75	5.25	2.00	NaN	NaN	NaN	NaN

In [ ]:

```
df_target.tail()
```

Out[ ]:

	Toán	Ngữ văn	Địa lí	Tiếng Anh	Sinh học	Vật lí	Hóa học	Lịch sử
34821	0.75	5.00	NaN	NaN	5.2	NaN	4.8	NaN
34822	4.75	5.75	NaN	3.33	4.6	6.8	4.6	NaN
34823	4.00	5.50	NaN	3.60	NaN	5.6	NaN	NaN
34824	5.75	6.00	NaN	2.88	NaN	7.4	4.8	NaN
34825	2.50	4.25	NaN	3.00	4.4	4.2	4.4	NaN

In [ ]:

```
df_result = pd.concat([df,df_target], axis=1)
```

In [ ]:

```
df_result.head()
```

Out[ ]:

	SOBAODANH	HO_TEN	NGAY_SINH	TEN_CUMTHI	GIOI_TINH	DIEM_THI	Toán	Ngữ văn	Đ
0	018000001	DƯƠNG VIỆT AN	12/03/1998	Sở GDĐT Bắc Giang	Nam	Toán: 2.00 Ngữ văn: 5.50 Lịch sử: 3....	2.00	5.50	5.(
1	018000002	ĐỖ VĂN AN	09/12/1998	Sở GDĐT Bắc Giang	Nam	Toán: 5.50 Ngữ văn: 5.25 Địa lí: 5.5...	5.50	5.25	5.5
2	018000003	ĐỖ XUÂN AN	12/08/1997	Sở GDĐT Bắc Giang	Nam	Toán: 4.50 Ngữ văn: 5.50 Địa lí: 3.7...	4.50	5.50	3.7
3	018000004	ĐẶNG PHÚC AN	19/03/1998	Sở GDĐT Bắc Giang	Nữ	Toán: 3.00 Ngữ văn: 6.00 Địa lí: 5.5...	3.00	6.00	5.5
4	018000005	ĐẶNG VĂN AN	25/10/1998	Sở GDĐT Bắc Giang	Nam	Toán: 2.25 Ngữ văn: 4.75 Địa lí: 5.2...	2.25	4.75	5.2

In [ ]:

```
df_result.shape
```

Out[ ]:

```
(34826, 14)
```

In [ ]:

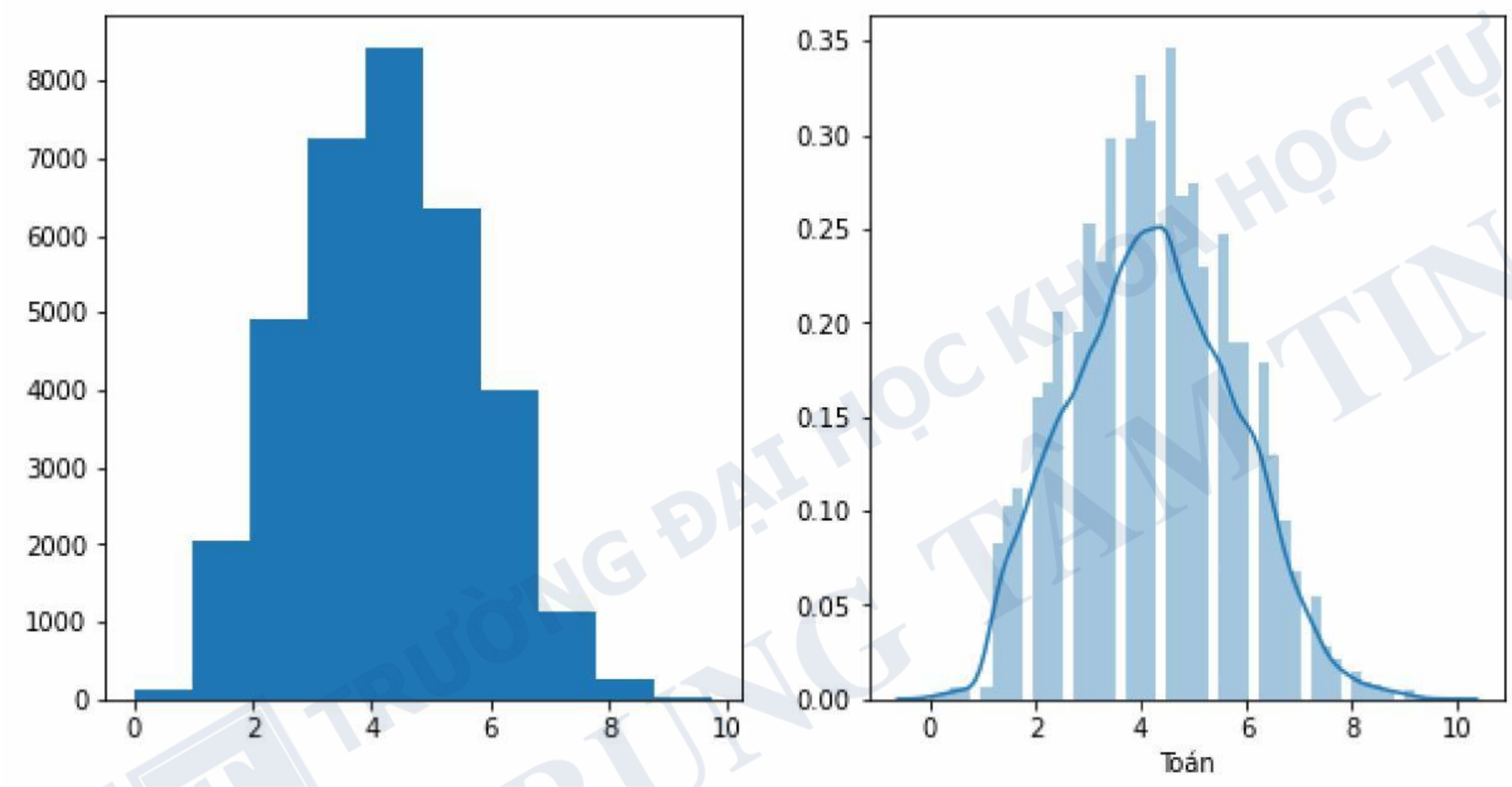
```
# Lưu dữ liệu sau khi đã chuẩn hóa
df_result.to_excel(path +"Diem_thi_THPT_2016_new_1.xlsx")
```

In [ ]:

```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:

```
plt.figure(figsize=(10, 5))
plt.subplot(1, 2, 1)
plt.hist(df_result['Toán'].dropna())
plt.subplot(1, 2, 2)
sns.distplot(df_result['Toán'].dropna())
plt.show()
```



In [ ]:

```
# HV thực hiện trực quan hóa và nhận xét cho các môn thi
```