Customer Segmentation - 고객 군집

```
import warnings
warnings.filterwarnings('ignore')
```

I. Import Packages and Google Drive Mount

1) Import Packages

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
```

2) Google Drive Mount

- 'Online_Retail.zip' 파일을 구글드라이브에 업로드 후 진행

```
from google.colab import drive

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

```
Mounted at /content/drive
```

- 마운트 결과 확인

```
!ls -l '/content/drive/My Drive/Colab Notebooks/datasets/Online_Retail.zip'
```

```
-rw------- 1 root root 22824989 Mar  7 07:09 '/content/drive/My Drive/Colab Notebooks/dataset
```

II. Data Preprocessing

1) Unzip 'Online_Retail.zip'

- Colab 파일시스템에 'Online_Retail.csv' 파일 생성

```
!unzip /content/drive/My Drive/Colab Notebooks/datasets/Online_Retail.zip
```

```
Archive:  /content/drive/My Drive/Colab Notebooks/datasets/Online_Retail.zip
  inflating: Online_Retail.xlsx
```

- Online_Retail.zip 파일 확인

```
!ls -l
```

```
total 23168
drwx------ 5 root root     4096 Mar 10 05:40 drive
-rw-r--r-- 1 root root 23715344 Mar  7 16:08 Online_Retail.xlsx
drwxr-xr-x 1 root root     4096 Mar  5 14:37 sample_data
```

2) 데이터 읽어오기

- pandas DataFrame

```
%%time

DF = pd.read_excel('Online_Retail.xlsx')

DF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype         
---  ------       --------------   -----         
 0   InvoiceNo    541909 non-null  object        
 1   StockCode    541909 non-null  object        
 2   Description  540455 non-null  object        
 3   Quantity     541909 non-null  int64         
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64       
 6   CustomerID   406829 non-null  float64       
 7   Country      541909 non-null  object        
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
CPU times: user 38.5 s, sys: 392 ms, total: 38.9 s
Wall time: 38.9 s
```

3) 데이터 설명

- InvoiceNo : 주문 번호, 'C' 시작은 주문취소
- StockCode : 제품 코드(Item Code)
- Description : 제품 설명
- Quantity : 주문 건수
- InvoiceDate : 주문 날짜
- UnitPrice : 제품 단가

- CustomerID : 고객번호
- Country : 국가명(주문 고객 국적)

```
DF.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	Cus
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	
2	536365	84406B	CREAM CUPID HEARTS COAT	8	2010-12-01 08:26:00	2.75	

4) 결측치 제거

- 'Quantity', 'UnitPrice', 'CustomerID'

```
DF = DF[DF['Quantity'] > 0]
DF = DF[DF['UnitPrice'] > 0]
DF = DF[DF['CustomerID'].notnull()]

DF.shape
```

```
(397884, 8)
```

- 결과 확인

```
DF.isnull().sum(axis = 0)
```

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

5) 'United Kingdom' 만 사용

- 대부분의 구매자가 영국국적

```
DF['Country'].value_counts()[:10]
```

```
United Kingdom    354321
```

```
Germany             9040
France              8341
EIRE                7236
Spain               2484
Netherlands         2359
Belgium             2031
Switzerland         1841
Portugal            1462
Australia           1182
Name: Country, dtype: int64
```

- 영국 데이터만 추출

```
DF = DF[DF['Country'] == 'United Kingdom']

DF.shape
```

```
(354321, 8)
```

#

#

#

The End

#

#

#