데이터 불러오기

cd라는 명령어는 현재 자기의 파일 위치(디렉토리, directory)를 알려주는 명령어

In [2]:

```
cd
```

C:₩Users₩an-user

cd 다음에 자기가 원하는 디렉토리를 설정하는 방법(뒤에 주소는 그냥 복붙)

In [3]:

```
cd C:₩Users₩an-user₩Desktop₩예비 데이터₩graduate-admissions
```

C:₩Users₩an-user₩Desktop₩예비 데이터₩graduate-admissions

In [4]:

```
##기본적으로 csv나 excel같은 것을 읽고 조작할 때 필요한 패키지 2개. pandas와 numpy를 불러옴
import pandas as pd ##앞으로 pandas 패키지를 이용할 때 pd라고 줄여서 쓸 예정
import numpy as np ## 앞으로 numpy 패키지를 이용할 때 np라고 줄여서 쓸 예정
```

In [5]:

```
##데이터를 읽어오는 방법 pd.read_csv('원하는 파일.csv')

df = pd.read_csv('Admission_Predict_Ver1.1.csv')
```

이제부터 데이터를 탐구해봅시다. (EDA - Exploratory Data Analysis) 쉽게 말해서 내가 가지고 있는 데이터가 어떻게 생겼는지, 구성이 되어있는지 간단하게 아이디어만 얻는 것.

In [6]:

```
##head라는 명령어는 모든 변수(columns)의 첫번째 5개만 보여주는 명령어
##주로 데이터셋이 잘 들어왔는지 확인하기 위해서 씀. 반대로 tail이라고 하면 끝에 5개를 보여줌
df.head()
```

Out[6]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65

In [7]:

```
##info()라는 명령어는 observation의 개수, column들의 data type, column들의 종류를
##보여주는 아주 유용한 명령어. 항상 처음에 이것을 쓰는 것 같음.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 9 columns):
Serial No.           500 non-null int64
GRE Score            500 non-null int64
TOEFL Score          500 non-null int64
University Rating    500 non-null int64
SOP                  500 non-null float64
LOR                  500 non-null float64
CGPA                 500 non-null float64
Research             500 non-null int64
Chance of Admit      500 non-null float64
dtypes: float64(4), int64(5)
memory usage: 35.2 KB
```

data type에 대한 지식은 필수라서 이 링크로 들어가서 표에 대해서 읽어보는 것을 추천드림
https://pbpython.com/pandas_dtypes.html (https://pbpython.com/pandas_dtypes.html)

In [8]:

```
##describe()라는 명령어는 변수의 평균, 중앙값, 표준편차 등을 보여주는 아주 유용한 명령어임.
##때에 따라서 아주 유용하게 많이 쓰임.
df.describe()
```

Out[8]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	C
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.00000	500.0(
mean	250.500000	316.472000	107.192000	3.114000	3.374000	3.48400	8.576
std	144.481833	11.295148	6.081868	1.143512	0.991004	0.92545	0.604
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.00000	6.800(
25%	125.750000	308.000000	103.000000	2.000000	2.500000	3.00000	8.127
50%	250.500000	317.000000	107.000000	3.000000	3.500000	3.50000	8.560(
75%	375.250000	325.000000	112.000000	4.000000	4.000000	4.00000	9.040(
max	500.000000	340.000000	120.000000	5.000000	5.000000	5.00000	9.920(

데이터에 대해서 이상한 점 살펴보기

1. serial number가 정말 float(정수)일까? - 이것은 identification을 위한 정보 따라서 알맞는 데이터 종류는 object.(Serial number is more like a name of a person rather than a value that could be calculated.)
2. data min or max 값에서 이상한 값들은 없는지. 만약 TOFEL 점수가 120점이 넘어간다면 그것은 데이터로서 가치가 있을까? 혹은 SOP 점수에서 혼자 100점이 나와 있다면 괜찮은 것일까?? 이러한 질문들에 대해서 답변해보고 이상한 값이 있는 column이 있다면 나중에 수정해야 할 것. 다행이도 여기서는 그러한 극단적인 수치는 없는 것 같음

이런 식으로 데이터에 대해서 기본적으로 이상한 점이 없는지 찾아보는 것이 시작 단계.

여기서 조금더 진도를 나가고 싶다면 그래프를 그려도 됨. 참고 자료로 내 프로젝트 따로 올리겠음.