토익진단평가 데이터 다듬기2

단계2 : 데이터프레임 탐색

0.미션

- 전처리 단계에서 생성한 데이터에 대한 데이터프레임 탐색을 해봅시다.
- 개별 변수 및 개별 변수들 간의 관계에 대해 생각해보고, 분석해봅니다.

1.환경설정

(1) 폰트 설치

In [373…
```
# 아래 라이브러리를 수행해주세요.
```

In [374…
```
!pip install matplotlib
!pip install --upgrade matplotlib

import matplotlib.pyplot as plt
```

```
Requirement already satisfied: matplotlib in c:\users\user\anaconda3\lib\site-packages (3.8.3)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\user\anaconda3\lib\site-packages
(from matplotlib) (1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\users\user\anaconda3\lib\site-packages (from
matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\user\anaconda3\lib\site-packages
(from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\user\anaconda3\lib\site-packages
(from matplotlib) (1.4.4)
Requirement already satisfied: numpy<2,>=1.21 in c:\users\user\anaconda3\lib\site-packages (fr
om matplotlib) (1.24.3)
Requirement already satisfied: packaging>=20.0 in c:\users\user\anaconda3\lib\site-packages (f
rom matplotlib) (23.1)
Requirement already satisfied: pillow>=8 in c:\users\user\anaconda3\lib\site-packages (from ma
tplotlib) (10.0.1)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\user\anaconda3\lib\site-packages
(from matplotlib) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\user\anaconda3\lib\site-packag
es (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\user\anaconda3\lib\site-packages (from pyt
hon-dateutil>=2.7->matplotlib) (1.16.0)
Requirement already satisfied: matplotlib in c:\users\user\anaconda3\lib\site-packages (3.8.3)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\user\anaconda3\lib\site-packages
(from matplotlib) (1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\users\user\anaconda3\lib\site-packages (from
matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\user\anaconda3\lib\site-packages
(from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\user\anaconda3\lib\site-packages
(from matplotlib) (1.4.4)
Requirement already satisfied: numpy<2,>=1.21 in c:\users\user\anaconda3\lib\site-packages (fr
om matplotlib) (1.24.3)
Requirement already satisfied: packaging>=20.0 in c:\users\user\anaconda3\lib\site-packages (f
rom matplotlib) (23.1)
Requirement already satisfied: pillow>=8 in c:\users\user\anaconda3\lib\site-packages (from ma
tplotlib) (10.0.1)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\user\anaconda3\lib\site-packages
(from matplotlib) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\user\anaconda3\lib\site-packag
es (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\user\anaconda3\lib\site-packages (from pyt
hon-dateutil>=2.7->matplotlib) (1.16.0)
```

(2) 라이브러리 불러오기

- **세부 요구사항**
  - 기본적으로 필요한 라이브러리를 import 하도록 코드를 작성하였습니다.
  - pandas, numpy, matplotlib 라이브러리를 실행해주세요.

In [375…
```
#[문제] pandas 라이브러리를 임포트하세요.
```

In [376…
```
import pandas as pd
```

In [377…
```
#[문제] numpy 라이브러리를 임포트하세요.
```

In [378…
```
import numpy as np
```

In [379…
```
#[문제] matplotlib 라이브러리를 임포트하세요.
```

In [380…
```
import matplotlib.pyplot as plt
```

In [381…
```
#차트에 한글폰트 설정을 위해 아래 라이브러리를 실행해주세요.
```

In [382…
```
plt.rc('font', family='Malgun Gothic')
```

(3) 데이터 불러오기

- 1.데이터 전처리 단계에서 앞 시간에 생성한 데이터를 로딩합니다.
  - data04_baseline.csv
- 다음과 같이 데이터를 저장하여 불러와주세요.
  - 주피터랩 수행
    - 제공된 압축파일 '미프 1차_토익'을 다운받아 압축을 푼다.
    - anaconda의 root directory(보통 C:\Users\ 에 '미프 1차_토익' 폴더를 만들고, 복사해 넣습니다.
    - '2.데이터_탐색_교육생용' 실습파일을 열어주세요.

1) 주피터랩 수행

In [383…
```
# '미프 1차_토익' 폴더에 필요한 파일들을 넣고, 본 파일을 열고 데이터를 읽어옵니다.
```

In [384…
```
#[문제] '미프 1차_토익' 폴더에서 본 파일 '2.데이터_탐색_교육생용' 실습파일을 열어주세요.
```

In [ ]:
```
```

2) 저장된 데이터 읽어오기

- [1.전처리] 단계에서 저장된 .csv 파일인 'data04_baseline.csv' 파일을 불러옵니다.
- 불러 온 후에는 shape를 확인해 봅시다.

In [385…
```
#[문제] data04_baseline.csv 파일을 pd.read_csv 함수를 이용하여 읽고 data변수에 할당하세요.
```

In [386…
```
# 읽어들일 파일명 : data04_baseline.csv
# Pandas read_csv 함수 활용
# 결과 : data 저장
```

In [387…
```
data = pd.read_csv('data04_baseline_1_csv')
```

In [388…
```
#[문제] 읽어온 데이터프레임 data 확인해주세요.
```

In [389…
```
data
```

Out[389]:

	ID	Seq	3st_LC_Score	3st_RC_Score	3st_Total_Score	학습목표	학습방법	강의학습교재유형	학습빈도	기출문제공부횟수	...	토익모의테스트횟수	1st_LC_Sc
0	1	3	345	336	681	승진	온라인강의	영상교재	주5-6회	7.0	...	10	1
1	2	3	380	368	748	승진	온라인강의	뉴스/이슈기반교재	주5-6회	4.0	...	14	3
2	3	3	416	382	798	자기계발	참고서	일반적인영어텍스트기반교재	주1-2회	4.0	...	4	3
3	4	3	495	397	892	승진	온라인강의	뉴스/이슈기반교재	주3-4회	9.0	...	8	4
4	5	3	398	437	835	자기계발	온라인강의	영상교재	주3-4회	6.0	...	4	2
...	...	...	...	...	...	...	...	...	...	...	...	...	
495	496	3	364	336	700	자기계발	온라인강의	일반적인영어텍스트기반교재	매일(주7회)	10.0	...	13	3
496	497	3	187	252	439	승진	온라인강의	비즈니스시뮬레이션(Role Play)	매일(주7회)	9.0	...	17	1
497	498	3	255	167	422	자기	오프라	일반적인영어	주1-	0.0	...	4	2

	ID	Seq	3st_LC_Score	3st_RC_Score	3st_Total_Score	학습목표	학습방법	강의학습교재유형	학습빈도	기출문제공부횟수	...	토익모의테스트횟수	1st_LC_Sc
						계발	인강의	텍스트기반교재	2회				
498	499	3	422	370	792	자기계발	오프라인강의	비즈니스시뮬레이션(Role Play)	주3-4회	4.0	...	7	3
499	500	3	235	226	461	승진	오프라인강의	비즈니스시뮬레이션(Role Play)	주5-6회	7.0	...	15	1

In [390…
```
#[문제] data 데이터프레임의 열과 행을 확인해주세요.
```

In [391…
```
data.shape
```

Out[391]:
```
(500, 21)
```

In [392…
```
#[문제] data 데이터프레임의 자료구조(Row, Column, Not-null, type)을 파악하세요.
```

In [393…
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                500 non-null    int64
 1   Seq               500 non-null    int64
 2   3st_LC_Score      500 non-null    int64
 3   3st_RC_Score      500 non-null    int64
 4   3st_Total_Score   500 non-null    int64
 5   학습목표              500 non-null    object
 6   학습방법              500 non-null    object
 7   강의 학습 교재 유형       500 non-null    object
 8   학습빈도              500 non-null    object
 9   기출문제 공부 횟수        500 non-null    float64
 10  취약분야 인지 여부        500 non-null    object
 11  토익 모의테스트 횟수       500 non-null    int64
 12  1st_LC_Score      500 non-null    int64
 13  1st_RC_Score      500 non-null    int64
 14  1st_Total_Score   500 non-null    int64
 15  2st_LC_Score      500 non-null    int64
 16  2st_RC_Score      500 non-null    int64
 17  2st_Total_Score   500 non-null    int64
 18  Gender            500 non-null    object
 19  Birth_Year        500 non-null    int64
 20  Score_diff_total  500 non-null    int64
dtypes: float64(1), int64(14), object(6)
memory usage: 82.2+ KB
```

In [394…
```
#[문제] 인덱스를 확인해보세요.
```

In [395…
```
data.index
```

Out[395]:
```
RangeIndex(start=0, stop=500, step=1)
```

In [396…
```
#[문제] 컬럼명을 확인해보세요.
```

In [397…
```
data.columns
```

Out[397]:
```
Index(['ID', 'Seq', '3st_LC_Score', '3st_RC_Score', '3st_Total_Score', '학습목표',
       '학습방법', '강의 학습 교재 유형', '학습빈도', '기출문제 공부 횟수', '취약분야 인지 여
부',
       '토익 모의테스트 횟수', '1st_LC_Score', '1st_RC_Score', '1st_Total_Score',
       '2st_LC_Score', '2st_RC_Score', '2st_Total_Score', 'Gender',
       'Birth_Year', 'Score_diff_total'],
      dtype='object')
```

In [398…
```
#[문제] 상단 5행을 확인해보세요
```

In [399…
```
data.head()
```

Out[399]:

	ID	Seq	3st_LC_Score	3st_RC_Score	3st_Total_Score	학습목표	학습방법	강의학습교재유형	학습빈도	기출문제공부횟수	...	토익모의테스트횟수	1st_LC_Score	1s
0	1	3	345	336	681	승진	온라인강의	영상교재	주5-6회	7.0	...	10	181	
1	2	3	380	368	748	승진	온라인강의	뉴스/이슈기반교재	주5-6회	4.0	...	14	330	
2	3	3	416	382	798	자기계발	참고서	일반적인영어텍스트기반교재	주1-2회	4.0	...	4	367	
3	4	3	495	397	892	승진	온라인강의	뉴스/이슈기반교재	주3-4회	9.0	...	8	470	
4	5	3	398	437	835	자기계발	온라인강의	영상교재	주3-4회	6.0	...	4	273	

5 rows × 21 columns

2.데이터프레임 탐색 : 개별 변수 분석하기

- **세부 요구사항**
  - 기본 분석
    - 기초 통계량, NaN 값 확인 등 기본 분석을 수행합니다.
  - 주요 변수들의 분포를 살펴보기
    - 숫자형 변수 : 기초통계량 조회 (평균, 중앙값, 표준편차 등)
    - 범주형 변수 : 특정한 카테고리나 범주로 구성된 변수인 범주별 빈도수, 바 플롯 (예시 : 성별, 혈액형, 지역 등 그룹화할 수 있는 변수)
    - 시각화나 통계분석에서 범주형 변수의 그룹 간의 차이나 관계를 분석할 때 중요한 역할을 합니다.

(1) 기본 분석

- **세부 요구사항**
  - 데이터프레임 전체에 대한 기초통계량을 구합니다.
  - NaN을 확인해봅니다.
    - NaN이 존재한다면 Numpy 모듈에서 isna() 함수를 통해 판별하고 적절하게 조치해 줍니다.

In [400…
```
#[문제] 'data'의 각 열별 누락된 값(Nan, none) 개수를 확인해보세요.
```

In [401…
```
#isna() 함수 활용
data.isna().sum()
```

Out[401]:
```
ID                  0
Seq                 0
3st_LC_Score        0
3st_RC_Score        0
3st_Total_Score     0
학습목표                0
학습방법                0
강의 학습 교재 유형         0
학습빈도                0
기출문제 공부 횟수          0
취약분야 인지 여부          0
토익 모의테스트 횟수         0
1st_LC_Score        0
1st_RC_Score        0
1st_Total_Score     0
2st_LC_Score        0
2st_RC_Score        0
2st_Total_Score     0
Gender              0
Birth_Year          0
Score_diff_total    0
dtype: int64
```

In [402…
```
#[문제] 'data'의 각 열 통계량을 요약하여 출력하세요.
```

In [403…
```
# describe 함수 활용
data.describe()
```

Out[403]:

	ID	Seq	3st_LC_Score	3st_RC_Score	3st_Total_Score	기출문제 공부 횟수	토익 모의 테스트 횟수	1st_LC_S
count	500.000000	500.0	500.000000	500.000000	500.000000	500.000000	500.000000	500.00
mean	250.500000	3.0	368.240000	369.518000	737.798000	5.092000	9.460000	313.87
std	144.481833	0.0	82.135393	81.665858	155.901584	2.789103	4.955554	85.55
min	1.000000	3.0	141.000000	135.000000	280.000000	0.000000	1.000000	105.00
25%	125.750000	3.0	295.000000	295.000000	591.750000	3.000000	5.000000	259.75
50%	250.500000	3.0	372.500000	375.000000	760.500000	5.000000	8.000000	308.00
75%	375.250000	3.0	434.250000	437.250000	860.250000	7.000000	13.250000	369.25
max	500.000000	3.0	495.000000	495.000000	990.000000	10.000000	20.000000	495.00

In [404…
```
#[문제] 위에서 출력한 통계량 데이터프레임의 행과 열을 서로 바꾸어 출력하세요.
```

In [405…
```
#'T'는 데이터프레임의 전치(Transpose)를 의미함
# descibe 함수 활용
data.describe().T
```

Out[405]:

	count	mean	std	min	25%	50%	75%	max
ID	500.0	250.500	144.481833	1.0	125.75	250.5	375.25	500.0
Seq	500.0	3.000	0.000000	3.0	3.00	3.0	3.00	3.0
3st_LC_Score	500.0	368.240	82.135393	141.0	295.00	372.5	434.25	495.0
3st_RC_Score	500.0	369.518	81.665858	135.0	295.00	375.0	437.25	495.0
3st_Total_Score	500.0	737.798	155.901584	280.0	591.75	760.5	860.25	990.0
기출문제 공부 횟수	500.0	5.092	2.789103	0.0	3.00	5.0	7.00	10.0
토익 모의테스트 횟수	500.0	9.460	4.955554	1.0	5.00	8.0	13.25	20.0
1st_LC_Score	500.0	313.878	85.555611	105.0	259.75	308.0	369.25	495.0
1st_RC_Score	500.0	312.822	86.574966	84.0	250.00	311.5	377.25	491.0
1st_Total_Score	500.0	626.700	148.571710	250.0	519.00	642.0	735.00	970.0
2st_LC_Score	500.0	338.120	84.169535	120.0	279.00	333.5	395.25	495.0
2st_RC_Score	500.0	338.154	83.854382	129.0	281.50	335.0	400.00	495.0
2st_Total_Score	500.0	676.284	153.178624	260.0	557.75	691.0	790.50	990.0
Birth_Year	500.0	1992.906	8.224381	1973.0	1986.75	1992.5	2000.00	2007.0
Score_diff_total	500.0	61.514	39.739051	0.0	30.00	63.0	83.00	281.0

In [406…
```
#[문제] Gender 컬럼의 값 별 개수를 확인해주세요. (Gender M : 남자, F : 여자)
```

In [407…
```
# value_counts 함수 활용
data['Gender'].value_counts()
```

Out[407]:
```
Gender
M    250
F    250
Name: count, dtype: int64
```

In [408…
```
#[문제] 'Gender' 컬럼의 ['M', 'F'] --> [1,2]로 변경해해보세요.
```

In [409…
```
# replace 함수 활용
data['Gender'] = data['Gender'].replace({'M': 1, 'F': 2})
```

In [410…
```
#[문제] Gender 컬럼의 값 별 개수를 다시 확인해주세요.
```

In [411…
```
data['Gender'].value_counts()
```

Out[411]:
```
Gender
1    250
2    250
Name: count, dtype: int64
```

In [412…
```
#[문제] 'Gender' 컬럼 타입을 object에서 int로 변경해보세요.
```

In [413…
```
#astype 함수 활용
data['Gender'] = data['Gender'].astype(int)
```

In [414…
```
data['Gender']
```

Out[414]:
```
0      1
1      2
2      2
3      1
4      1
      ..
495    1
496    2
497    1
498    2
499    1
Name: Gender, Length: 500, dtype: int32
```

In [415…
```
#[문제] data 데이터 프레임에서 Null 데이터가 있는지 확인해주세요.
```

In [416…
```
data.isna().sum()
```

Out[416]:
```
ID                  0
Seq                 0
3st_LC_Score        0
3st_RC_Score        0
3st_Total_Score     0
학습목표                0
학습방법                0
강의 학습 교재 유형         0
학습빈도                0
기출문제 공부 횟수          0
취약분야 인지 여부          0
토익 모의테스트 횟수         0
1st_LC_Score        0
1st_RC_Score        0
1st_Total_Score     0
2st_LC_Score        0
2st_RC_Score        0
2st_Total_Score     0
Gender              0
Birth_Year          0
Score_diff_total    0
dtype: int64
```

(2) 주요 변수의 분포를 살펴보기

- **세부 요구사항**
  - 주요 변수들의 분포를 살펴보자.
    - 대상 : 최종 차수 점수 변화, 생년월일, 학습목표, 학습방법, 강의 학습 교재 유형
    - 도구 : 기초통계량, bar차트를 통한 데이터 분포 파악
  - 전체 현황과 각 변수별 현황을 비교해보고 결과를 저장하자.

1) 열 데이터 탐색 및 시각화 : 최종 차수 점수 변화(Score_diff_total)

In [417…
```
#[문제] 변수 Sdt에 문자열 'Score_diff_total'을 할당 하세요.
```

In [418…
```
Sdt = 'Score_diff_total'
```

In [419…
```
#[문제] 'data' 데이터프레임의 'Score_diff_total'열에 대한 기술 통계 정보를 데이터 프레임의 형태
```

In [420…
```
p_data = data['Score_diff_total'].describe() #Std 같은 결과
p_data = pd.DataFrame(p_data)
p_data
```

Out[420]:

	Score_diff_total
count	500.000000
mean	61.514000
std	39.739051
min	0.000000
25%	30.000000
50%	63.000000
75%	83.000000
max	281.000000

In [421…

```
# describe() 함수 활용
p_data = data[Sdt].describe()  #Score_diff_total 결과
p_data = pd.DataFrame(p_data)
p_data
```

Out[421]:

	Score_diff_total
count	500.000000
mean	61.514000
std	39.739051
min	0.000000
25%	30.000000
50%	63.000000
75%	83.000000
max	281.000000

In [422…

```
#[문제] 위에서 추출한 데이터의 'Score_diff_total'에 대해 행과 열을 변환하여 기술 통계 정보를 출
```

In [423…

```
# describe().T 함수 활용
p_data = data[Sdt]
p_data = pd.DataFrame(p_data)
p_data.describe().T
```

Out[423]:

	count	mean	std	min	25%	50%	75%	max
Score_diff_total	500.0	61.514	39.739051	0.0	30.0	63.0	83.0	281.0

2) 열 추가 : Bitrh_Year

- 출생연도로 부터 나이 변수를 추출해 봅시다.

In [424…

```
#[문제] 변수 BY에 문자열 'Birth_Year'을 할당하세요.
```

In [425…
```
BY = 'Birth_Year'
```

In [426…
```
#[문제] 'data' 데이터프레임의 'Birth_Year'열에 대한 기술 통계 정보를 출력해주세요.
#'Birth_Year'에 대해 행과 열을 변환하여 기술 통계 정보를 출력해주세요.
```

In [427…
```
# describe().T 함수 활용
b_data = data[BY]
b_data = pd.DataFrame(b_data)
b_data.describe().T
```

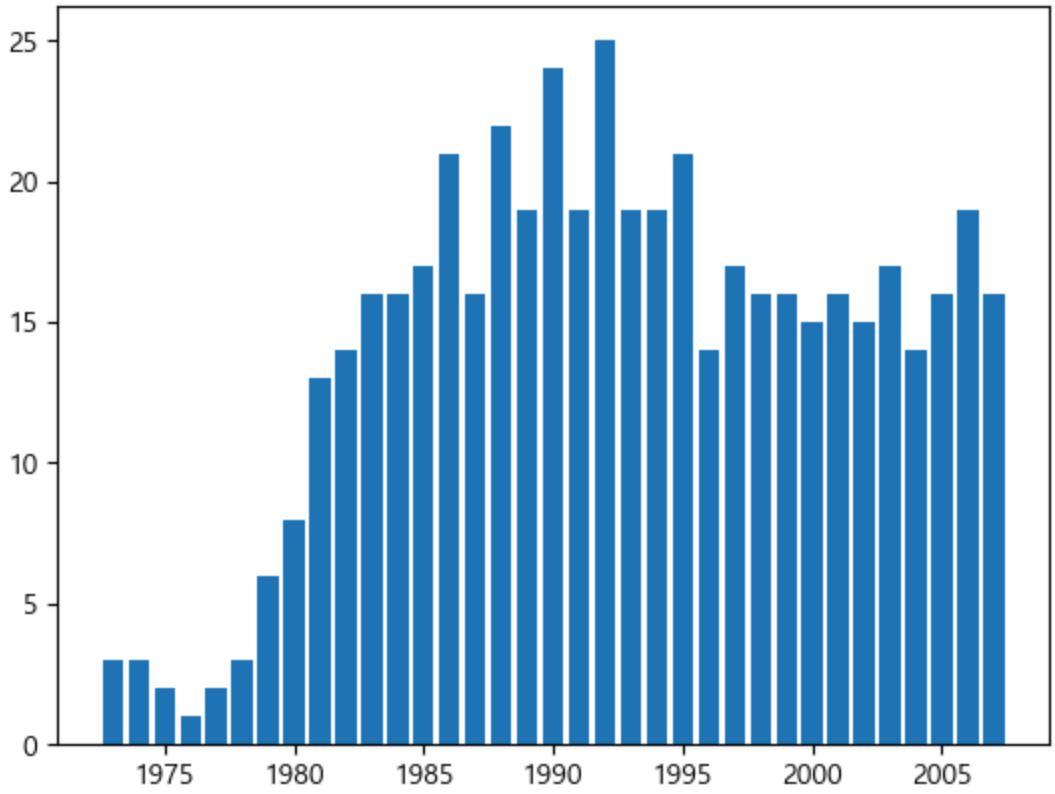
Out[427]:

	count	mean	std	min	25%	50%	75%	max
Birth_Year	500.0	1992.906	8.224381	1973.0	1986.75	1992.5	2000.0	2007.0

In [428…
```
#[문제] 'data' 데이터프레임의 'Birth_Year' 컬럼의 연도별 개수를 Bar 차트로 그리세요.
```

In [429…
```
# DataFrame value_counts()와 plot() 함수 활용
# 대상 컬럼 : 'Birth_Year'
# plot 함수의 인자 : kind='bar'

ppy = data.groupby(BY, as_index=False)['ID'].count()
#data[[BY]].value_counts().sort_index().plot(kind='bar')
plt.bar(ppy[BY], ppy['ID'])
plt.show()
```



3) 고유값 확인 : 컬럼 출력

In [430…
```
#[문제] data 데이터프레임에서 object 컬럼에 대해서만 추출해서 보여주세요.
```

In [431…
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                500 non-null    int64
 1   Seq               500 non-null    int64
 2   3st_LC_Score      500 non-null    int64
 3   3st_RC_Score      500 non-null    int64
 4   3st_Total_Score   500 non-null    int64
 5   학습목표              500 non-null    object
 6   학습방법              500 non-null    object
 7   강의 학습 교재 유형       500 non-null    object
 8   학습빈도              500 non-null    object
 9   기출문제 공부 횟수        500 non-null    float64
 10  취약분야 인지 여부        500 non-null    object
 11  토익 모의테스트 횟수       500 non-null    int64
 12  1st_LC_Score      500 non-null    int64
 13  1st_RC_Score      500 non-null    int64
 14  1st_Total_Score   500 non-null    int64
 15  2st_LC_Score      500 non-null    int64
 16  2st_RC_Score      500 non-null    int64
 17  2st_Total_Score   500 non-null    int64
 18  Gender            500 non-null    int32
 19  Birth_Year        500 non-null    int64
 20  Score_diff_total  500 non-null    int64
dtypes: float64(1), int32(1), int64(14), object(5)
memory usage: 80.2+ KB
```

In [432…
```
# select_dtypes() 함수 활용
data.select_dtypes('object')
```

Out[432]:

	학습목표	학습방법	강의 학습 교재 유형	학습빈도	취약분야 인지 여부
0	승진	온라인강의	영상 교재	주5-6회	알고 있음
1	승진	온라인강의	뉴스/이슈 기반 교재	주5-6회	알고 있음
2	자기계발	참고서	일반적인 영어 텍스트 기반 교재	주1-2회	알고 있음
3	승진	온라인강의	뉴스/이슈 기반 교재	주3-4회	알고 있음
4	자기계발	온라인강의	영상 교재	주3-4회	알고 있음
…	…	…	…	…	…
495	자기계발	온라인강의	일반적인 영어 텍스트 기반 교재	매일(주 7회)	알고 있음
496	승진	온라인강의	비즈니스 시뮬레이션(Role Play)	매일(주 7회)	알고 있음
497	자기계발	오프라인강의	일반적인 영어 텍스트 기반 교재	주1-2회	알고 있음
498	자기계발	오프라인강의	비즈니스 시뮬레이션(Role Play)	주3-4회	알고 있음
499	승진	오프라인강의	비즈니스 시뮬레이션(Role Play)	주5-6회	알고 있음

500 rows × 5 columns

In [433…
```
#[문제] 데이터 타입이 Object 형태인 컬럼의 컬럼명만 추출해서 출력해 보세요.
```

In [434…
```
# columns.values 활용
O_data = data.select_dtypes(include='object').columns.values
```

In [435…
```
O_data
```

Out[435]:
```
array(['학습목표', '학습방법', '강의 학습 교재 유형', '학습빈도', '취약분야 인지 여부'], dtype
=object)
```

4) 고유값 확인 및 시각화 : 학습목표

In [436…
```
#[문제] 변수 '학습목표'의 값들의 빈도수를 계산하여 출력하세요.
```

In [437…
```
# value_counts 함수 활용
data['학습목표'].value_counts()
```

Out[437]:
```
학습목표
자기계발    329
승진      155
취업       16
Name: count, dtype: int64
```

In [438…
```
#[문제] data 데이터프레임의 전체 열과 행 개수를 각각 출력해 보세요.
```

In [439…
```
#열의 개수
data.shape[1]
```

Out[439]:
```
21
```

In [440…
```
#행의 개수
data.shape[0]
```

Out[440]:
```
500
```

In [441…
```
#[문제] 변수 '학습목표'의 값들의 빈도수를 전체 데이터의 개수로 나누어서 해당 값들이 전체 데이E
```

In [442…
```
bin_data = data['학습목표'].value_counts() / data['학습목표'].value_counts().sum() * 100
```

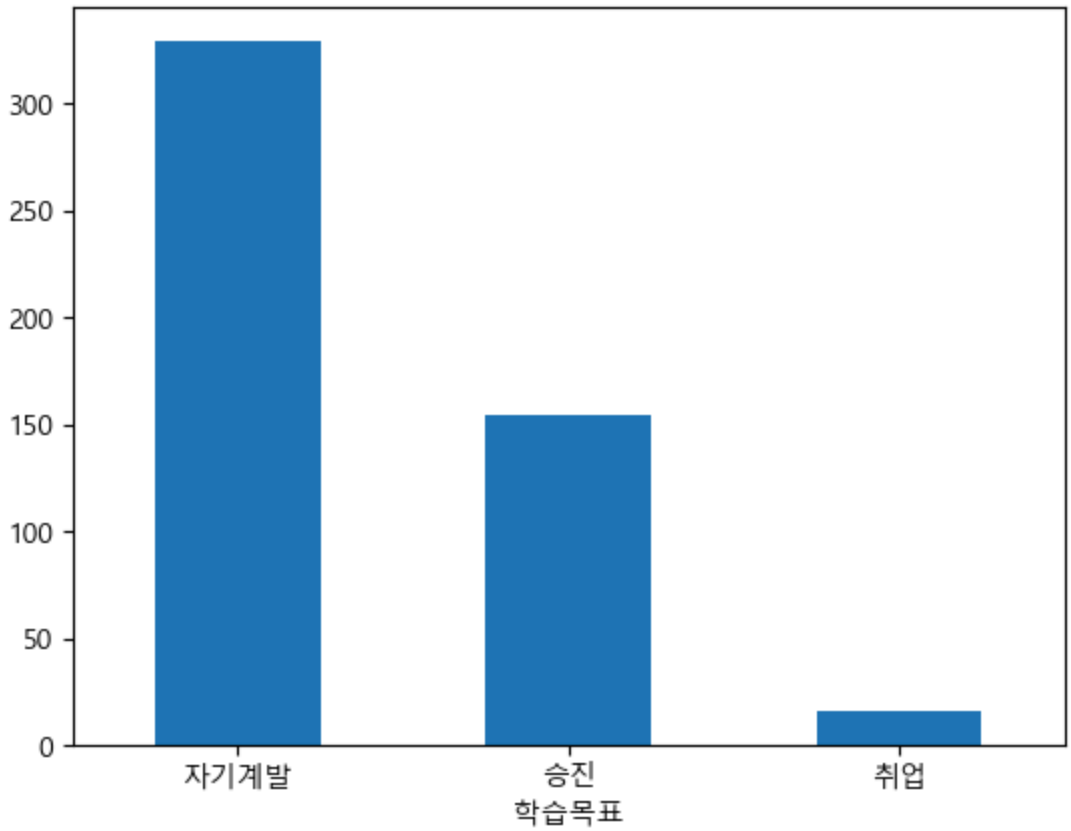
In [443…
```
bin_data
```

Out[443]:
```
학습목표
자기계발    65.8
승진      31.0
취업       3.2
Name: count, dtype: float64
```

In [444…
```
#[문제] '학습목표' 컬럼에 대한 Bar 차트를 확인해주세요.
```

In [445…
```
# DataFrame value_counts()와 plot() 함수 활용
# 대상 컬럼 : '학습목표'
# plot 함수의 인자 : kind='bar'
#goal_sdy =  data['학습목표'].value_counts()
#goal_sdy.plot(kind='bar')
data['학습목표'].value_counts().plot(kind='bar')
```

```
plt.xticks(rotation=0)
plt.show()
```



In [446…
```
#[문제] 'data' 전체 데이터프레임의 숫자형 컬럼, number 컬럼에 대해 검색해주세요.
```

In [447…
```
data.select_dtypes('number').columns
```

Out[447]:
```
Index(['ID', 'Seq', '3st_LC_Score', '3st_RC_Score', '3st_Total_Score',
       '기출문제 공부 횟수', '토익 모의테스트 횟수', '1st_LC_Score', '1st_RC_Score',
       '1st_Total_Score', '2st_LC_Score', '2st_RC_Score', '2st_Total_Score',
       'Gender', 'Birth_Year', 'Score_diff_total'],
      dtype='object')
```

5) 고유값 확인 및 시각화 : 강의 학습 교재 유형

In [448…
```
#[문제] 변수 '강의 학습 교재 유형'의 값들의 빈도수, 비율을 계산해서 출력해주세요.
```

In [449…
```
# 빈도수 출력
data['강의 학습 교재 유형'].value_counts()
```

Out[449]:
```
강의 학습 교재 유형
일반적인 영어 텍스트 기반 교재        136
영상 교재                      128
뉴스/이슈 기반 교재                122
비즈니스 시뮬레이션(Role Play)     114
Name: count, dtype: int64
```

In [450…
```
# 비율 출력
data['강의 학습 교재 유형'].value_counts() / data['강의 학습 교재 유형'].value_counts().sum()
```

Out[450]:
```
강의 학습 교재 유형
일반적인 영어 텍스트 기반 교재        27.2
영상 교재                    25.6
뉴스/이슈 기반 교재              24.4
비즈니스 시뮬레이션(Role Play)    22.8
Name: count, dtype: float64
```

In [451…
```
#[문제] '취약분야 인지 여부' 문자열의 값을 '알고 있음' --> '1', '알고 있지 않음' -->'0'으로 변
```

In [452…
```
#data = pd.get_dummies(data, columns=['취약분야 인지 여부'], drop_first=True, dtype=int)
data['취약분야 인지 여부'].replace({'알고 있음':1, '알고 있지 않음':0}, inplace=True)
```

In [453…
```
data['취약분야 인지 여부'].value_counts()
```

Out[453]:
```
취약분야 인지 여부
1    461
0     39
Name: count, dtype: int64
```

데이터 저장

- **세부 요구사항**
  - to_csv를 이용하여 전처리된 데이터셋을 저장하세요.
  - 저장할 파일의 확장자는 .csv 입니다.

In [454…
```
#[문제] 전처리된 데이터프레임 'data04_featured'를 CSV 파일로 저장합니다.
```

In [455…
```
# 파일 : 'data04_featured.csv'
# to_csv 함수 활용
data.to_csv('data04_featured_1.csv')
```

In [456…
```
#[문제] 파일이 잘 저장되었는지, 다시 한번 불러오고 확인해보세요.
```

In [457…
```
d_data = pd.read_csv('data04_featured_1.csv')
d_data
```

Out[457]:

	Unnamed: 0	ID	Seq	3st_LC_Score	3st_RC_Score	3st_Total_Score	학습목표	학습방법	강의학습교재유형	학습빈도	...	토익모의테스트횟수	1st_
0	0	1	3	345	336	681	승진	온라인강의	영상교재	주5-6회	...	10	
1	1	2	3	380	368	748	승진	온라인강의	뉴스/이슈기반교재	주5-6회	...	14	
2	2	3	3	416	382	798	자기계발	참고서	일반적인영어텍스트기반교재	주1-2회	...	4	
3	3	4	3	495	397	892	승진	온라인강의	뉴스/이슈기반교재	주3-4회	...	8	
4	4	5	3	398	437	835	자기계발	온라인강의	영상교재	주3-4회	...	4	
...	...	...	...	...	...	...	...	...	...	...	...	...	
495	495	496	3	364	336	700	자기계발	온라인강의	일반적인영어텍스트기반교재	매일(주7회)	...	13	
496	496	497	3	187	252	439	승진	온라인강의	비즈니스시뮬레이션(Role Play)	매일(주7회)	...	17	
497	497	498	3	255	167	422	자기	오프라	일반적인영어	주1-	...	4	

	Unnamed: 0	ID	Seq	3st_LC_Score	3st_RC_Score	3st_Total_Score	학습목표	학습방법	강의학습교재유형	학습빈도	...	토익모의테스트횟수	1st_
							계발	인강의	텍스트기반교재	2회			
498	498	499	3	422	370	792	자기계발	오프라인강의	비즈니스시뮬레이션(Role Play)	주3-4회	...	7	
499	499	500	3	235	226	461	승진	오프라인강의	비즈니스시뮬레이션(Role Play)	주5-6회	...	15	

In [ ]:
```
## 고생 정말 많으셨습니다!!
## 실습시간이 남으신 분은 '중급'용 파일에 도전해보세요.
```