상관분석
EDA

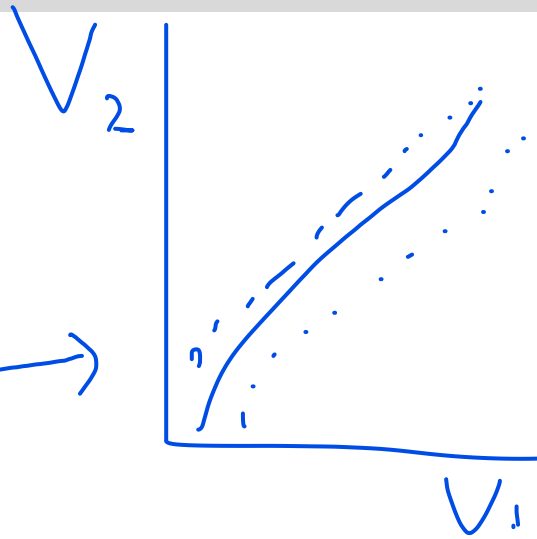
빅 데 이 터 아 카 데 미

상관 분석과 EDA1

1. 상관분석 개요
2. 상관계수
3. EDA 개요
4. EDA Case Study

상관 분석(Correlation Analysis)

- 상관 분석(Correlation Analysis)
  - 상관분석은 연속형 변수로 측정된 두 변수 간의 선형적 관계를 분석
  - 선형적 관계는 비례식이 성립되는 관계
  - A 변수가 증가함에 따라 B 변수도 증가되는지 혹은 감소하는지를 분석
  - 상관분석에서 두 변수 사이의 선형적인 관계 정도를 나타내기 위해 상관계수(correlation coefficient)를 사용



$y = \alpha + \beta x$

$\beta x + \alpha$

$-1 < \quad < 1$

기울기

절편값

상관 분석(Correlation Analysis)

▪ 공분산(Covariance)

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

두 확률변수의 분포가 결합된 결합확률분포의 분산
두 확률변수의 상관 방향성을 나타내는 척도

- x값이 x의 평균보다 클 때, y값이 y의 평균보다 크면 Cov(X, Y)는 양수(정비례)
- x값이 x의 평균보다 클 때, y값이 y의 평균보다 작으면 Cov(X, Y)는 음수(반비례)
- x의 값과 y의 값이 독립관계일때, Cov(X, Y)는 0

상관 정도의 척도로서는 유용하지 않음
상관 정도의 척도는 공분산을 이용한 상관계수를 많이 사용

상관 계수(Correlation Coefficient)

- 피어슨 상관계수(Pearson Correlation Coefficient)

Pearson 상관계수는 공분산 Cov를 이용하여 -1 ~ 1의 상관 정도를 나타내는 척도

$$\rho_{X,Y} = \frac{Cov(X,\ Y)}{\sqrt{V(X)V(Y)}}$$

- 상관계수가 1에 가까워질수록 강한 양의 상관관계
- 상관계수가 -1에 가까워질수록 강한 음의 상관관계
- 상관계수가 0에 가까워질수록 상관관계가 없음

상관 계수(Correlation Coefficient)

- 피어슨 상관계수(Pearson Correlation Coefficient)

상관계수	정도	관계
1	완전	양
0.7 ~ 0.9	강한	양
0.4 ~ 0.6	보통	양
0.1 ~ 0.3	약한	양
0	상관관계 미존재	
-0.1 ~ -0.3	약한	음
-0.4 ~ -0.6	보통	음
-0.7 ~ -0.9	강한	음
-1	완전	음

상관 분석 도구

- Iris 데이터 셋 상관관계 분석

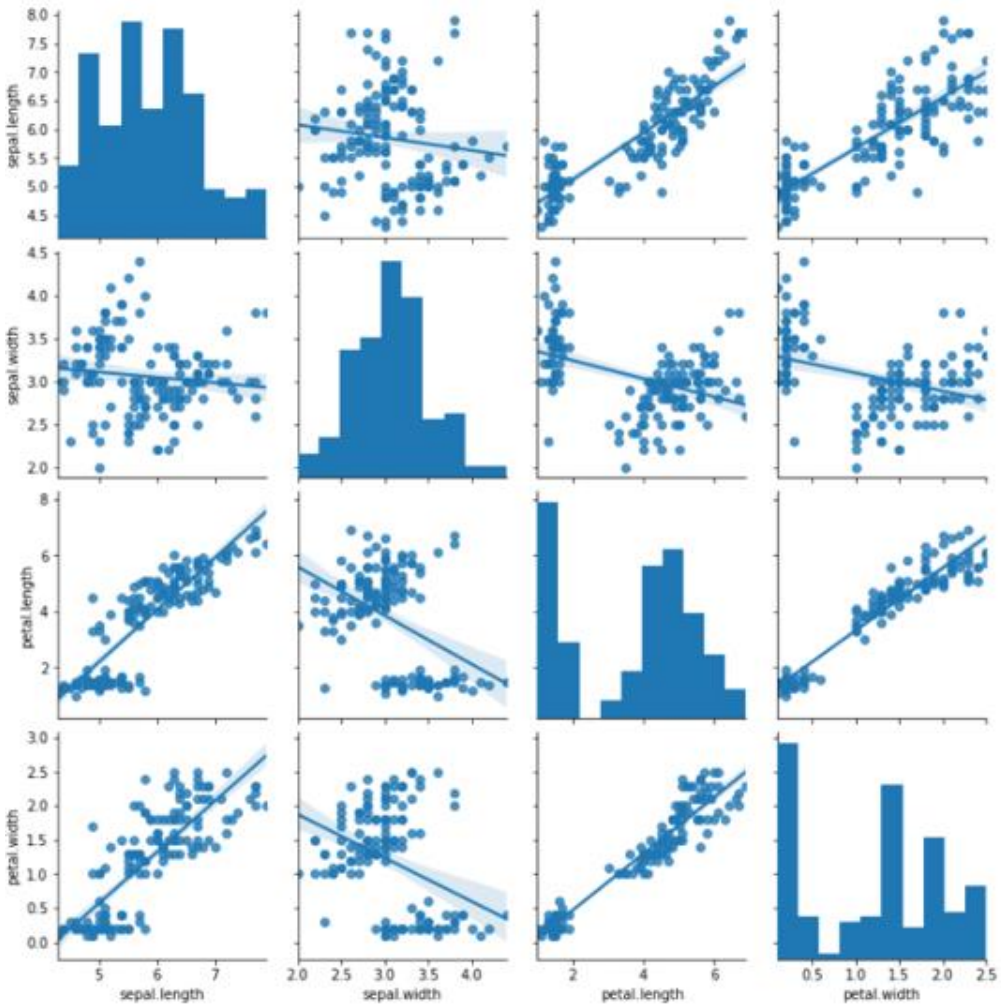
  Iris 데이터 셋 상관계수 행렬(Matrix)

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.117570	0.871754	0.817941
sepal_width	-0.117570	1.000000	-0.428440	-0.366126
petal_length	0.871754	-0.428440	1.000000	0.962865
petal_width	0.817941	-0.366126	0.962865	1.000000

상관 분석 도구

- Iris 데이터 셋 상관관계 분석

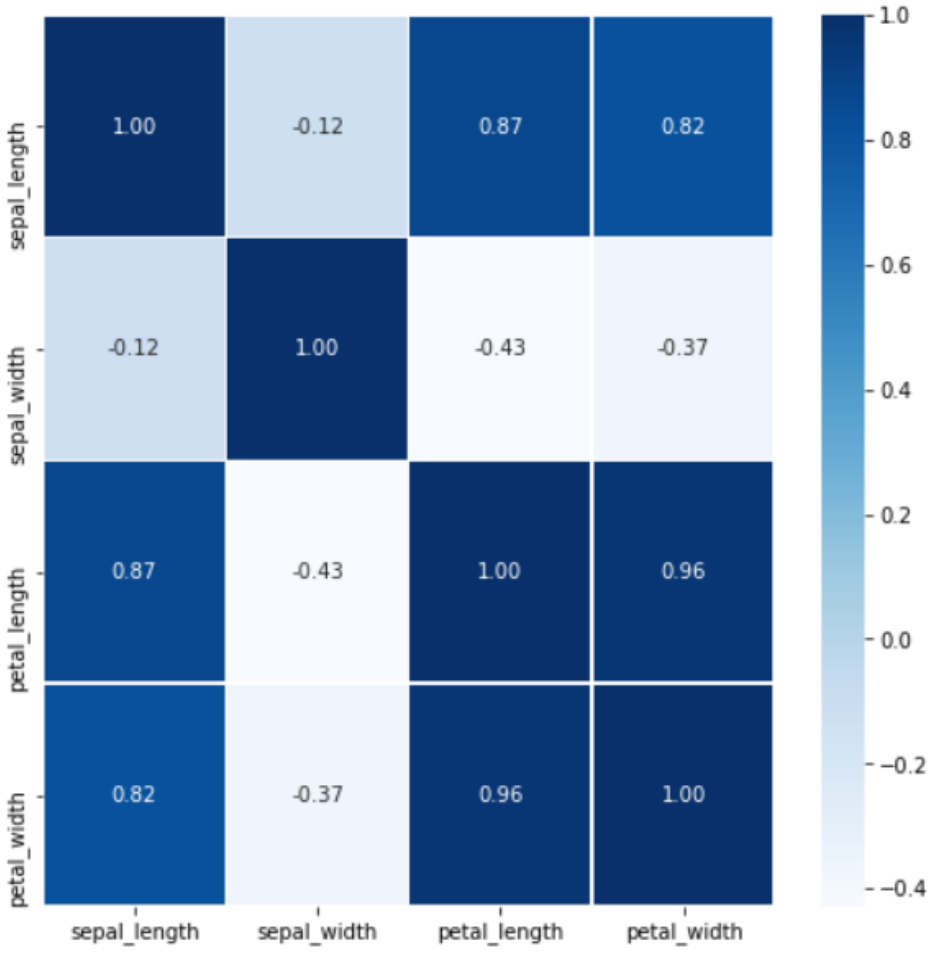
  Iris 데이터 셋 pair plot



상관 분석 도구

- Iris 데이터 셋 상관관계 분석

  Iris 데이터 셋 heat map



06-1. correlation.ipynb 실습

탐색적 데이터 분석(Exploratory Data Analysis)

- 탐색적 데이터 분석(Exploratory Data Analysis)이란?

  존 튜키(John W. Tukey) 의 저서『 탐색적 자료 분석(EDA) 』를 통해서 제시한 데이터 분석 방법론

  기술 통계학(descriptive statistics)의 중요성을 강조

  데이터가 가지고 있는 본연의 특징과 의미를 탐색

  데이터를 다양한 각도에서 관찰하고 이해

  데이터를 분석하기 전에 통계적인 방법이나 시각화 도구를 활용하여 데이터를 직관적으로 파악

- EDA의 목적

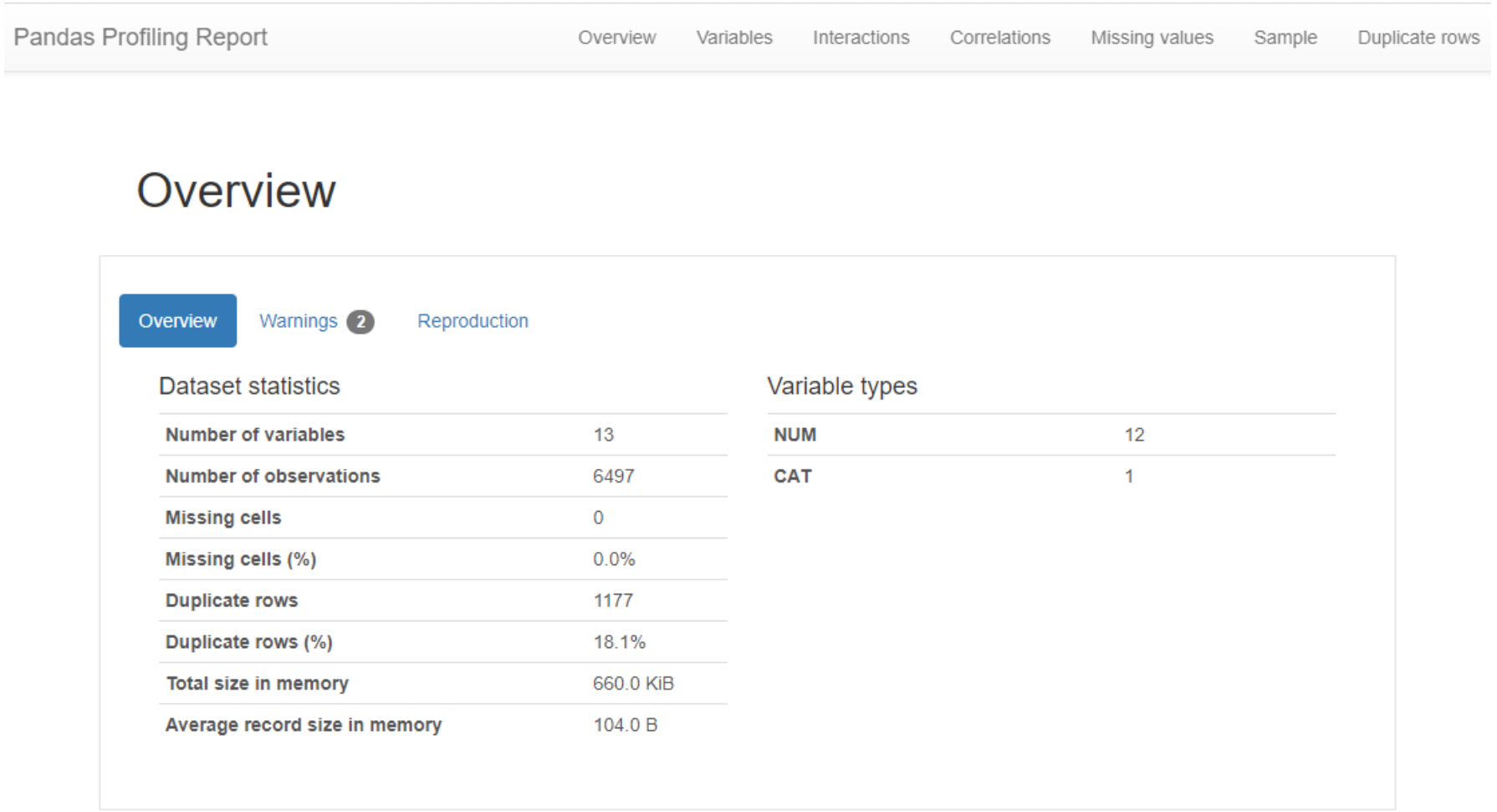
  데이터의 패턴을 파악하고 잠재적인 변수 간 관계를 이해

  이상치 또는 비정상적인 관측치와 같은 예외적 현상(anomalies)을 발견

  정형화된 통계 방법을 사용하여 검정할 수 있는 가설 수립을 위한 질문 도출

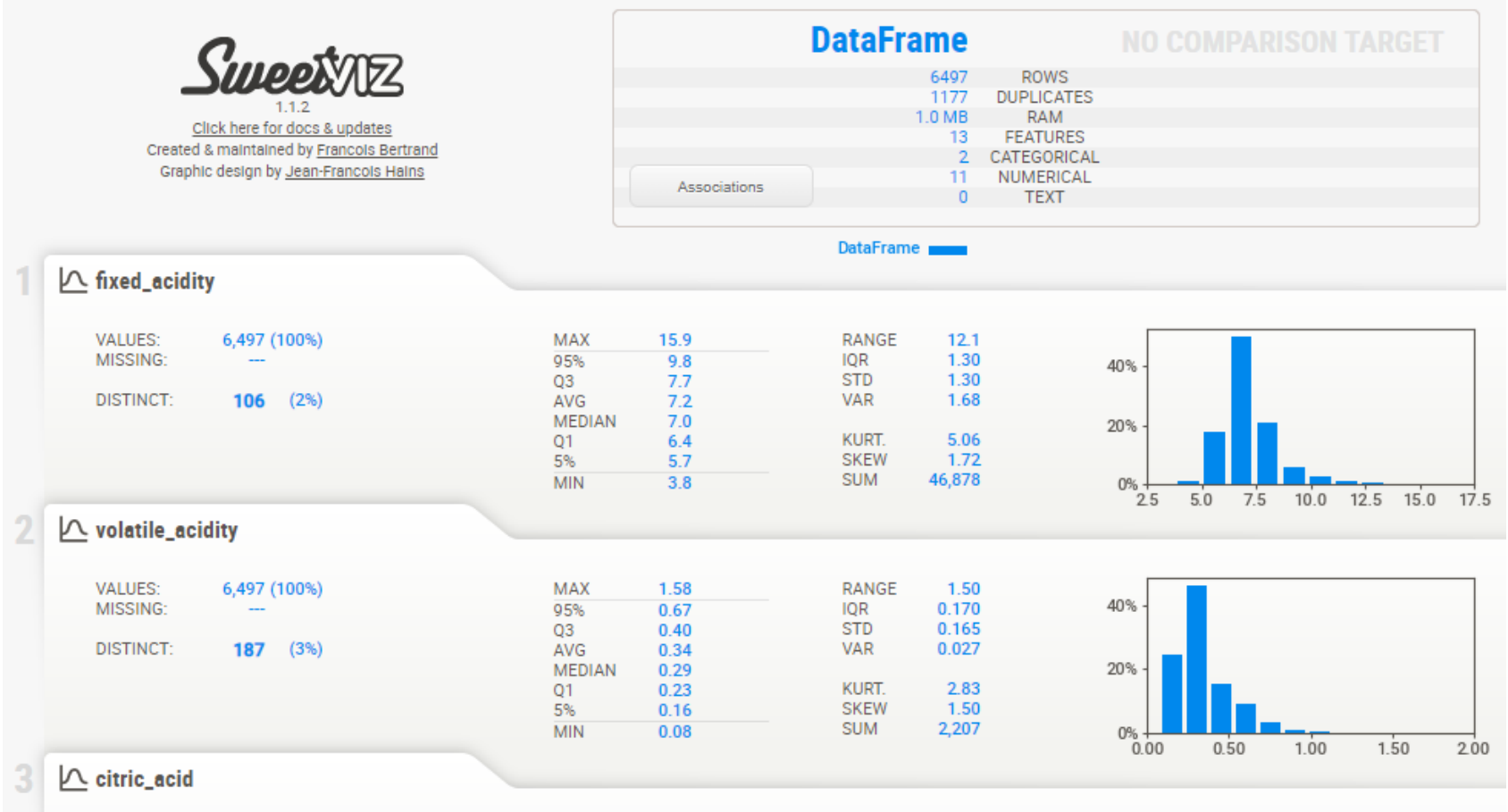
EDA 자동화 도구

- EDA 자동화 도구 – pandas data profiling



EDA 자동화 도구

- EDA 자동화 도구 – Sweetviz



EDA – CASE STUDY(와인 품질 데이터)

- 관측값 : 총 6,497건 (레드 와인: 1,599건, 화이트 와인: 4,898건)

- 입력변수 : 12개 (고정산, 휘발산, 구연산, 잔여당, 염화물, 무수아황산, 총이산화황, 밀도, 산성도, 황산염, 알콜도수와 같은 와인의 물리화학적 특성들과 red, white의 와인 타입)
- 출력변수 : 1개 (와인품질평가점수, 가장 낮은 품질 1점 ~ 가장 높은 품질 10점)

- 데이터 소스
  - https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv
  - https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv

EDA – 와인 품질 데이터

- 데이터 파일 read 및 데이터프레임 구성 확인

```
wine = pd.read_csv('wine.csv')
```

```
wine.head()
```

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	type
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	red
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	red
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	red
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red

EDA – 와인 품질 데이터

- 데이터 파일 read 및 데이터프레임 구성 확인

```
wine = pd.read_csv('wine.csv')
```

```
wine.head()
```

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	type
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	red
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	red
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	red
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red

EDA – 와인 품질 데이터

- 데이터 파일 read 및 데이터프레임 구성 확인

```
wine = pd.read_csv('wine.csv')
```

```
wine.head()
```

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	type
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	red
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	red
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	red
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red

EDA – 와인 품질 데이터

- pandas data profiling을 이용한 EDA

```
import pandas_profiling
profile = wine.profile_report()
```

```
profile.to_file('wine_profile.html')
```

Overview

Overview | Warnings 2 | Reproduction

Dataset statistics

Number of variables	13
Number of observations	6497
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	1177
Duplicate rows (%)	18.1%
Total size in memory	660.0 KiB
Average record size in memory	104.0 B

Variable types

NUM	12
CAT	1

EDA – 와인 품질 데이터

- Duplicate rows

Duplicate rows

Most frequent

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	q
460	7.0	0.15	0.28	14.7	0.051	29.0	149.0	0.99792	2.96	0.39	9.0	7
622	7.3	0.19	0.27	13.9	0.057	45.0	155.0	0.99807	2.94	0.41	8.8	8
360	6.8	0.18	0.30	12.8	0.062	19.0	171.0	0.99808	3.00	0.52	9.0	7
661	7.4	0.16	0.30	13.7	0.056	33.0	168.0	0.99825	2.90	0.44	8.7	7
660	7.4	0.16	0.27	15.5	0.050	25.0	135.0	0.99840	2.90	0.43	8.7	7
664	7.4	0.19	0.30	12.8	0.053	48.5	229.0	0.99860	3.14	0.49	9.1	7
665	7.4	0.19	0.31	14.5	0.045	39.0	193.0	0.99860	3.10	0.50	9.2	6
728	7.6	0.20	0.30	14.2	0.056	53.0	212.5	0.99900	3.14	0.46	8.9	8
32	5.7	0.22	0.20	16.0	0.044	41.0	113.0	0.99862	3.22	0.46	8.9	6
118	6.2	0.23	0.36	17.2	0.039	37.0	130.0	0.99946	3.23	0.43	8.8	6

EDA – 와인 품질 데이터

- 중복데이터 제거

```
wine[wine.duplicated(wine.columns, keep='last')]
```

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	type
0	7.4	0.700	0.00	1.90	0.076	11.0	34.0	0.99780	3.51	0.56	9.400000	5	red
9	7.5	0.500	0.36	6.10	0.071	17.0	102.0	0.99780	3.35	0.80	10.500000	5	red
22	7.9	0.430	0.21	1.60	0.106	10.0	37.0	0.99660	3.17	0.91	9.500000	5	red
39	7.3	0.450	0.36	5.90	0.074	12.0	87.0	0.99780	3.33	0.83	10.500000	5	red
64	7.2	0.725	0.05	4.65	0.086	4.0	11.0	0.99620	3.41	0.39	10.900000	5	red
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6424	6.0	0.340	0.29	6.10	0.046	29.0	134.0	0.99462	3.48	0.57	10.700000	6	white
6447	7.0	0.360	0.35	2.50	0.048	67.0	161.0	0.99146	3.05	0.56	11.100000	6	white
6448	6.4	0.330	0.44	8.90	0.055	52.0	164.0	0.99488	3.10	0.48	9.600000	5	white
6454	7.1	0.230	0.39	13.70	0.058	26.0	172.0	0.99755	2.90	0.46	9.000000	6	white
6478	6.6	0.340	0.40	8.10	0.046	68.0	170.0	0.99494	3.15	0.50	9.533333	6	white

1177 rows × 13 columns

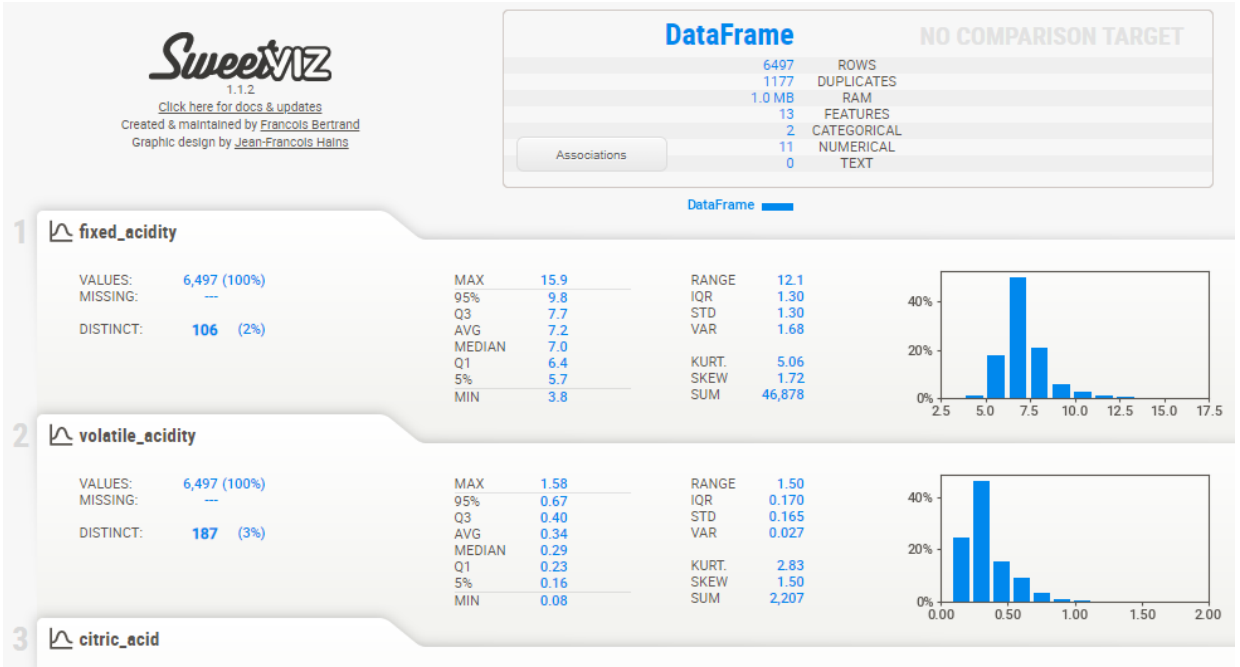
EDA – 와인 품질 데이터

- sweetviz를 이용한 EDA

```
!pip install sweetviz
```

```
import sweetviz
```

```
eda_report=sweetviz.analyze(wine)
eda_report.show_html()
```



EDA – 와인 품질 데이터

- 컬럼명 한글 변환 및 정보 출력

```
wine.columns = ['고정산', '휘발산', '구연산', '잔여당', '염화물', '무수아황산', '총이산화황',
                '밀도', '산성도', '황산염', '알콜도수', '와인품질', '와인종류']
```

```
wine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   고정산     6497 non-null   float64
 1   휘발산     6497 non-null   float64
 2   구연산     6497 non-null   float64
 3   잔여당     6497 non-null   float64
 4   염화물     6497 non-null   float64
 5   무수아황산   6497 non-null   float64
 6   총이산화황   6497 non-null   float64
 7   밀도      6497 non-null   float64
 8   산성도     6497 non-null   float64
 9   황산염     6497 non-null   float64
 10  알콜도수    6497 non-null   float64
 11  와인품질    6497 non-null   int64
 12  와인종류    6497 non-null   object
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

EDA – 와인 품질 데이터

- 연속형 데이터에 대한 기술 통계

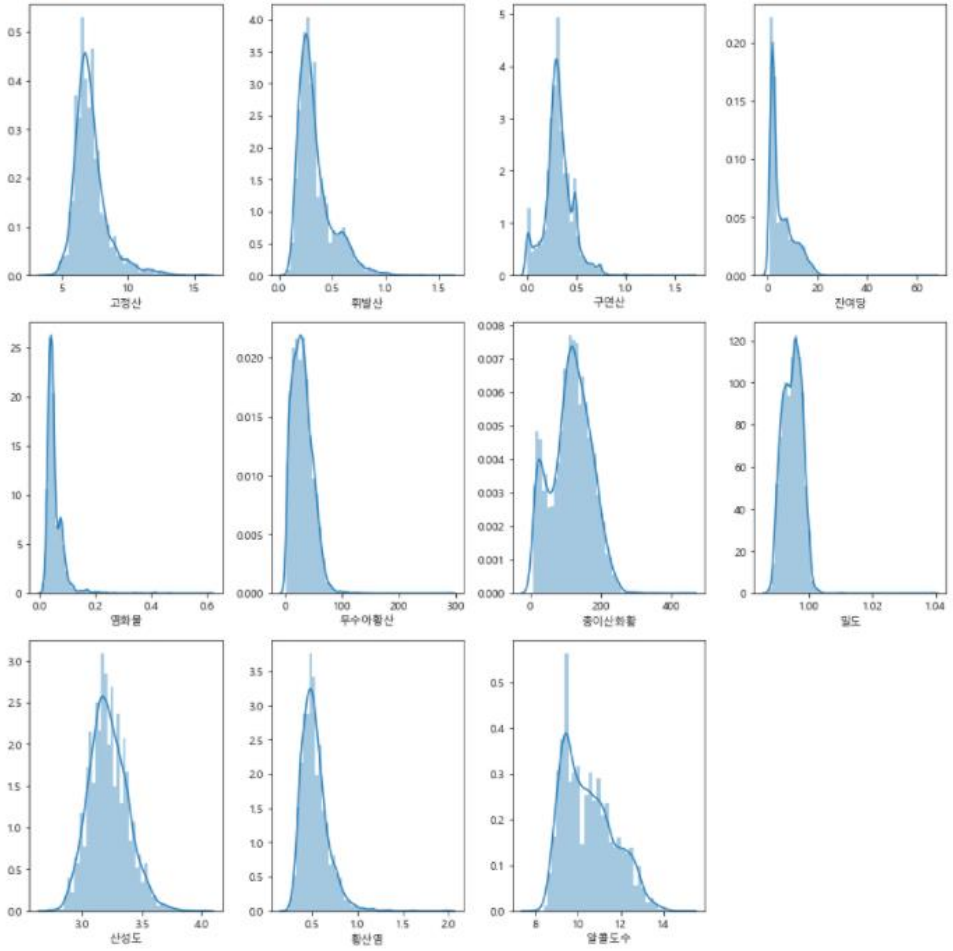
```
wine.describe()
```

	고정산	휘발산	구연산	잔여당	염화물	무수아황산	총이산화황	밀도	산성도	황산염	알콜도수	와인품질
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	115.744574	0.994697	3.218501	0.531268	10.491801	5.818378
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	56.521855	0.002999	0.160787	0.148806	1.192712	0.873255
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	0.220000	8.000000	3.000000
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	0.430000	9.500000	5.000000
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	0.510000	10.300000	6.000000
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	0.600000	11.300000	6.000000
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.010000	2.000000	14.900000	9.000000

EDA – 와인 품질 데이터

- 각 변수별 분포 subplot으로 시각화

```
plt.figure(figsize=(12,12))
for i in range(0,11):
    plt.subplot(3,4,i+1)
    sns.distplot(wine.iloc[:,i])
plt.tight_layout()
plt.show()
```



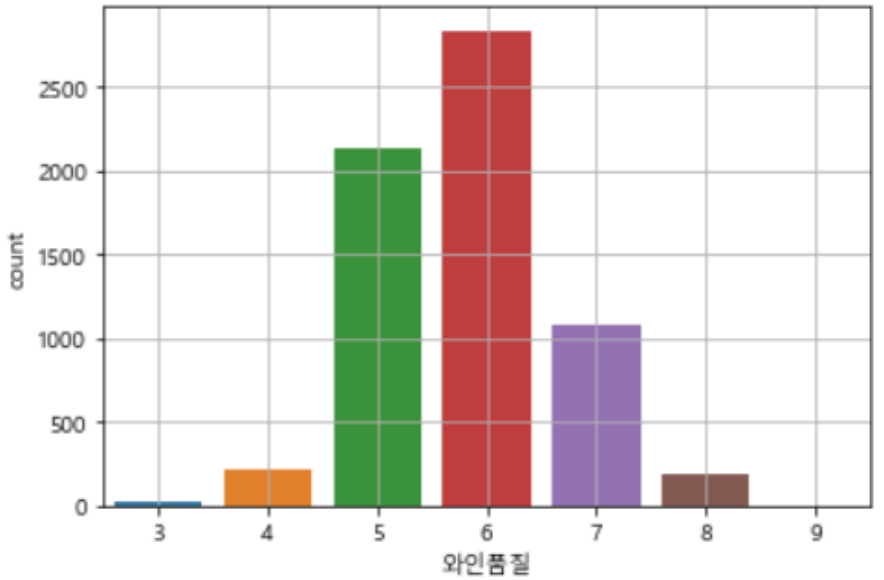
EDA – 와인 품질 데이터

- 와인 품질의 분포

```
pd.DataFrame(wine.와인품질.value_counts())
```

	와인품질
6	2836
5	2138
7	1079
4	216
8	193
3	30
9	5

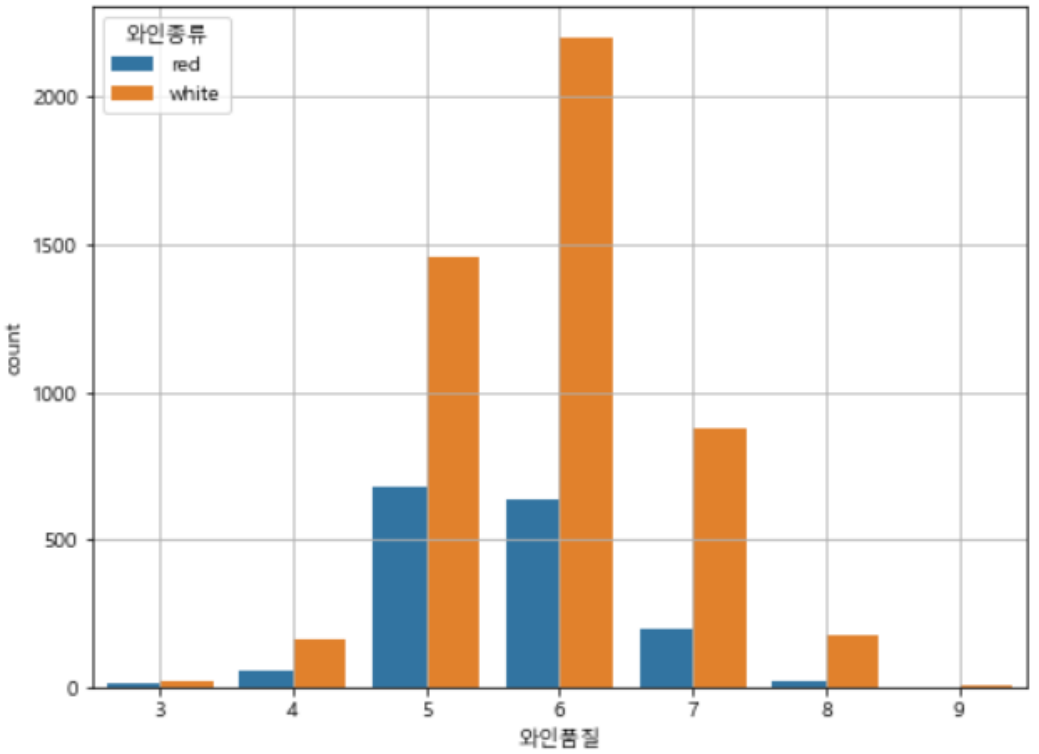
```
sns.countplot(x='와인품질', data=wine)
plt.grid()
plt.show()
```



EDA – 와인 품질 데이터

- 와인 종류별 와인 품질

```
plt.figure(figsize=(8, 6))
sns.countplot(x='와인품질',hue='와인종류', data=wine)
plt.grid()
plt.show()
```



EDA – 와인 품질 데이터

- 상관관계 행렬(Matrix)

```
wine_corr = wine.corr()
```

```
wine_corr
```

	고정산	휘발산	구연산	잔여당	염화물	무수아황산	총이산화황	밀도	산성도	황산염	알콜도수	와인품질
고정산	1.000000	0.219008	0.324436	-0.111981	0.298195	-0.282735	-0.329054	0.458910	-0.252700	0.299568	-0.095452	-0.076743
휘발산	0.219008	1.000000	-0.377981	-0.196011	0.377124	-0.352557	-0.414476	0.271296	0.261454	0.225984	-0.037640	-0.265699
구연산	0.324436	-0.377981	1.000000	0.142451	0.038998	0.133126	0.195242	0.096154	-0.329808	0.056197	-0.010493	0.085532
잔여당	-0.111981	-0.196011	0.142451	1.000000	-0.128940	0.402871	0.495482	0.552517	-0.267320	-0.185927	-0.359415	-0.036980
염화물	0.298195	0.377124	0.038998	-0.128940	1.000000	-0.195045	-0.279630	0.362615	0.044708	0.395593	-0.256916	-0.200666
무수아황산	-0.282735	-0.352557	0.133126	0.402871	-0.195045	1.000000	0.720934	0.025717	-0.145854	-0.188457	-0.179838	0.055463
총이산화황	-0.329054	-0.414476	0.195242	0.495482	-0.279630	0.720934	1.000000	0.032395	-0.238413	-0.275727	-0.265740	-0.041385
밀도	0.458910	0.271296	0.096154	0.552517	0.362615	0.025717	0.032395	1.000000	0.011686	0.259478	-0.686745	-0.305858
산성도	-0.252700	0.261454	-0.329808	-0.267320	0.044708	-0.145854	-0.238413	0.011686	1.000000	0.192123	0.121248	0.019506
황산염	0.299568	0.225984	0.056197	-0.185927	0.395593	-0.188457	-0.275727	0.259478	0.192123	1.000000	-0.003029	0.038485
알콜도수	-0.095452	-0.037640	-0.010493	-0.359415	-0.256916	-0.179838	-0.265740	-0.686745	0.121248	-0.003029	1.000000	0.444319
와인품질	-0.076743	-0.265699	0.085532	-0.036980	-0.200666	0.055463	-0.041385	-0.305858	0.019506	0.038485	0.444319	1.000000

EDA – 와인 품질 데이터

- 와인 품질에 대한 상관계수의 정렬

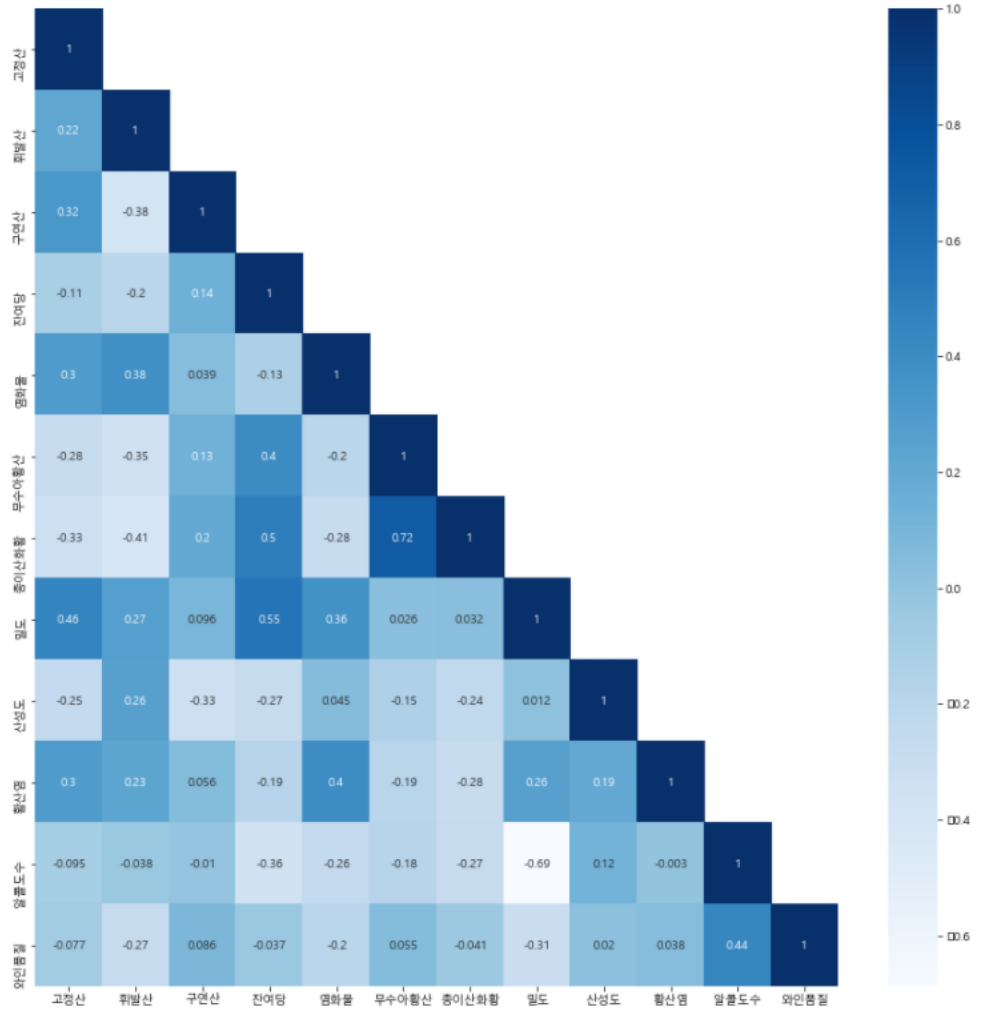
```
pd.DataFrame(wine_corr.와인품질.sort_values(ascending = False))
```

	와인품질	
와인품질	1.000000	
알콜도수	0.444319	양의 상관관계
구연산	0.085532	
무수아황산	0.055463	
황산염	0.038485	
산성도	0.019506	
잔여당	-0.036980	음의 상관관계
총이산화황	-0.041385	
고정산	-0.076743	
염화물	-0.200666	
휘발산	-0.265699	
밀도	-0.305858	

EDA – 와인 품질 데이터

- heatmap을 이용한 상관관계 시각화

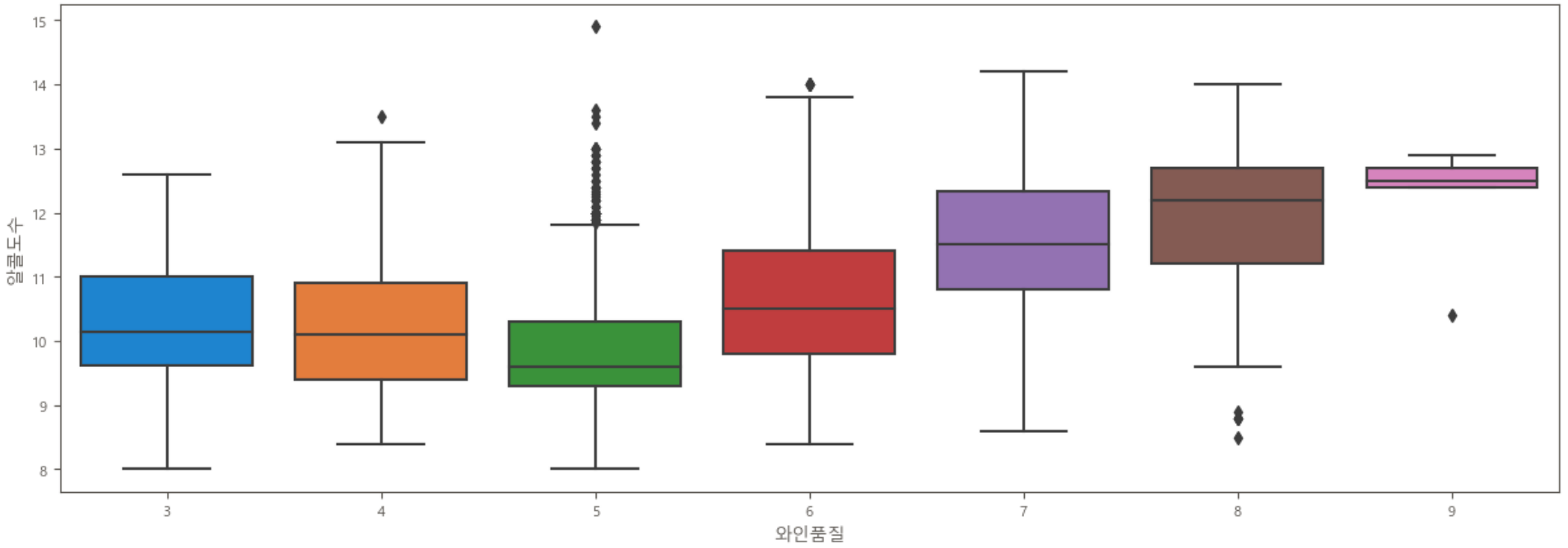
```
plt.figure(figsize=(15,15))
mask = np.array(wine_corr)
mask[np.tril_indices_from(mask)] = False
sns.heatmap(wine_corr, mask = mask, annot=True, cmap='Blues')
plt.show()
```



EDA – 와인 품질 데이터

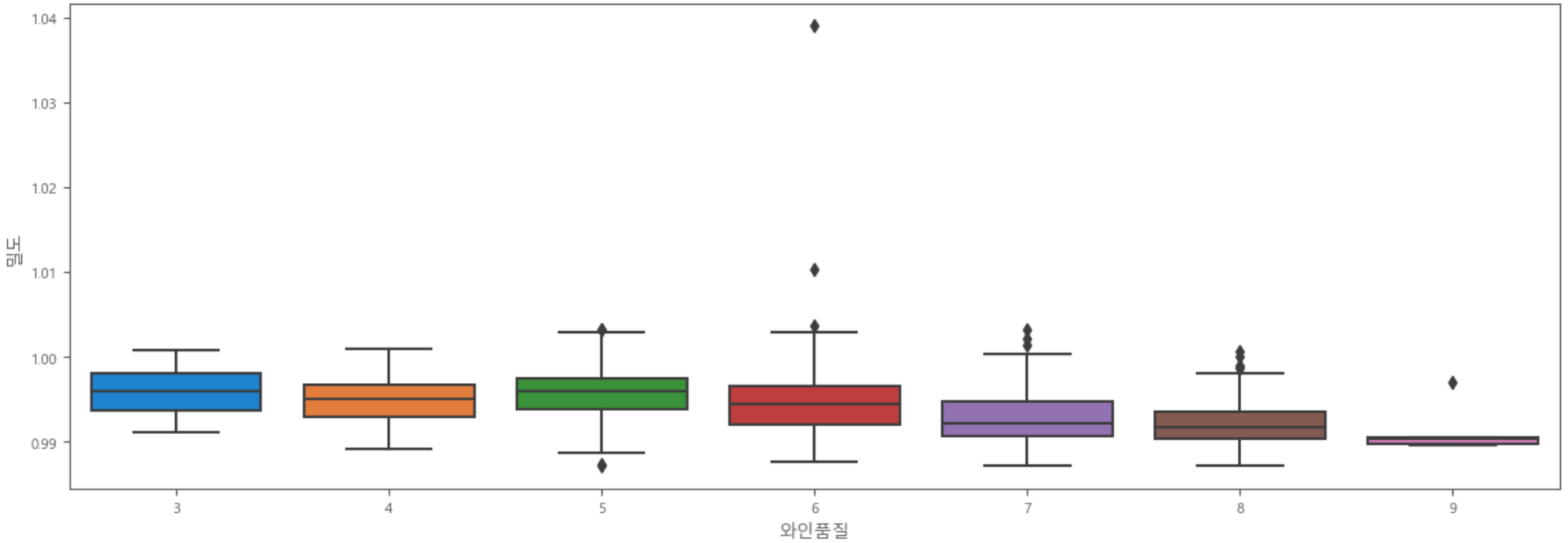
- 와인품질과 알콜 도수

```
plt.figure(figsize=(15,5))
sns.boxplot(x = '와인품질', y = '알콜도수', data = wine)
plt.show()
```



EDA – 와인 품질 데이터

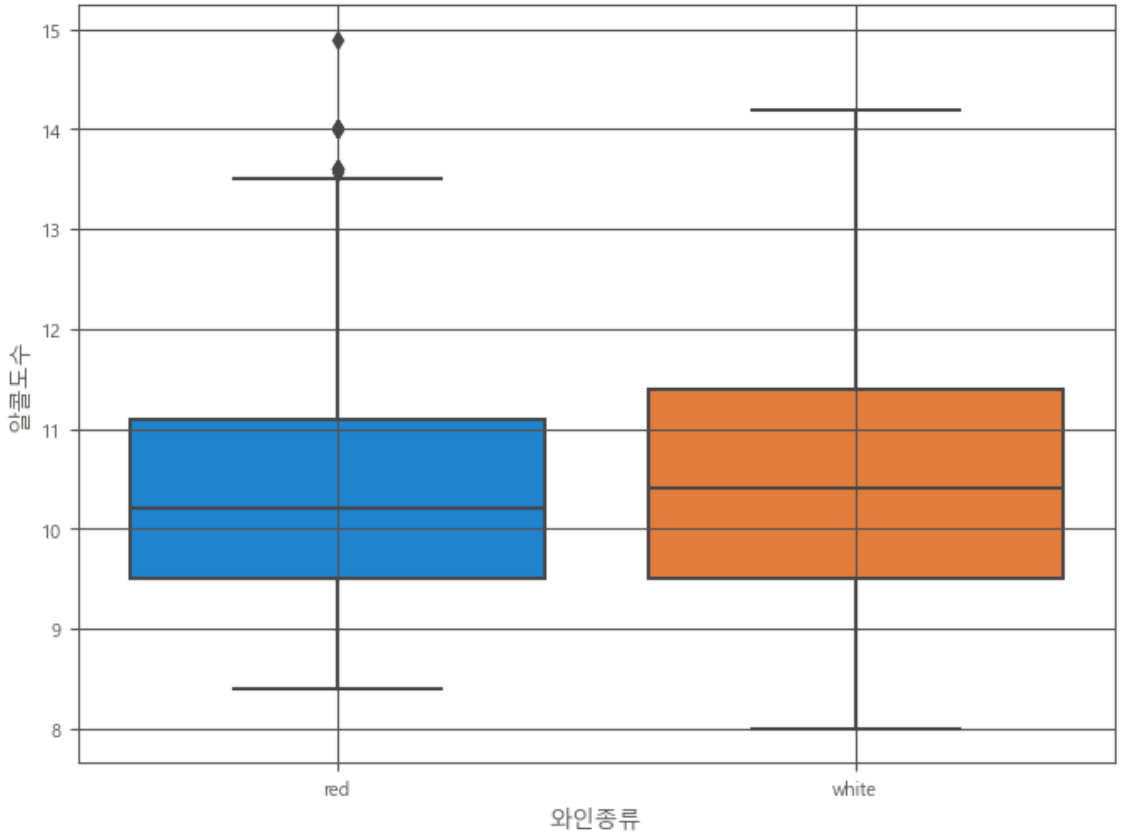
- 와인품질과 밀도



EDA – 와인 품질 데이터

- 와인종류와 알콜도수

```
plt.figure(figsize=(8, 6))
sns.boxplot(x = '와인종류', y = '알콜도수', data = wine)
plt.grid()
plt.show()
```



EDA – CASE STUDY

- 와인 종류에 따라 알콜 도수에 차이가 있을 까?

  – 귀무가설 : 와인 종류에 따라 알콜 도수에 차이가 없다.

  – 대립가설 : 와인 종류에 따라 알콜 도수에 차이가 있다.

```
red = wine[wine.와인종류 == 'red']['알콜도수']
white = wine[wine.와인종류 == 'white']['알콜도수']
```

```
tTestResult= stats.ttest_ind(red, white)
tTestResult
```

```
Ttest_indResult(statistic=-4.218888835968011, pvalue=2.4959339763303842e-05)
```

p-value = 2.49e-5 〈 0.05이므로

95% 신뢰수준하에서 귀무가설 기각, 대립가설 채택

와인종류에 따라 알콜 도수는 통계적으로 유의미한 차이가 있다.

06-2. eda_wine.ipynb 실습

32