```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 %matplotlib inline
5 import seaborn as sns
```

```
1 df = pd.read_csv('/content/winequalityN.csv')
```

```
1 df.head()
```

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	p
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.0
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.3
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.2
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.1
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.1

```
1 df.tail()
```

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density
6492	red	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490
6493	red	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512
6494	red	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574
6495	red	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547
6496	red	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   type                  6497 non-null   object 
 1   fixed acidity         6487 non-null   float64
 2   volatile acidity      6489 non-null   float64
 3   citric acid           6494 non-null   float64
 4   residual sugar        6495 non-null   float64
 5   chlorides             6495 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6488 non-null   float64
 10  sulphates             6493 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64  
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

```
1 df.corr()
```

```
<ipython-input-36-2f6f6606aa2c>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version
  df.corr()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	p
fixed acidity	1.000000	0.220172	0.323736	-0.112319	0.298421	-0.283317	-0.329747	0.459204	-0.25181
volatile acidity	0.220172	1.000000	-0.378061	-0.196702	0.377167	-0.353230	-0.414928	0.271193	0.26066
citric acid	0.323736	-0.378061	1.000000	0.142486	0.039315	0.133437	0.195218	0.096320	-0.32868
residual sugar	-0.112319	-0.196702	0.142486	1.000000	-0.128902	0.403439	0.495820	0.552498	-0.26705
chlorides	0.298421	0.377167	0.039315	-0.128902	1.000000	-0.195042	-0.279580	0.362594	0.04480
free sulfur dioxide	-0.283317	-0.353230	0.133437	0.403439	-0.195042	1.000000	0.720934	0.025717	-0.14519
total sulfur dioxide	-0.329747	-0.414928	0.195218	0.495820	-0.279580	0.720934	1.000000	0.032395	-0.23768
density	0.459204	0.271193	0.096320	0.552498	0.362594	0.025717	0.032395	1.000000	0.01192
pH	-0.251814	0.260660	-0.328689	-0.267050	0.044806	-0.145191	-0.237687	0.011920	1.00000
sulphates	0.300380	0.225476	0.057613	-0.185745	0.395332	-0.188489	-0.275381	0.259454	0.19124
alcohol	-0.095603	-0.038248	-0.010433	-0.359706	-0.256861	-0.179838	-0.265740	-0.686745	0.12100

```
1 df.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	su
count	6487.000000	6489.000000	6494.000000	6495.000000	6495.000000	6497.000000	6497.000000	6497.000000	6488.000000	649
mean	7.216579	0.339691	0.318722	5.444326	0.056042	30.525319	115.744574	0.994697	3.218395	
std	1.296750	0.164649	0.145265	4.758125	0.035036	17.749400	56.521855	0.002999	0.160748	
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.010000	2

```
1 df.duplicated()
```

```
0       False
1       False
2       False
3       False
4        True
        ...
6492    False
6493    False
6494     True
6495    False
6496    False
Length: 6497, dtype: bool
```

```
1 df.shape
```

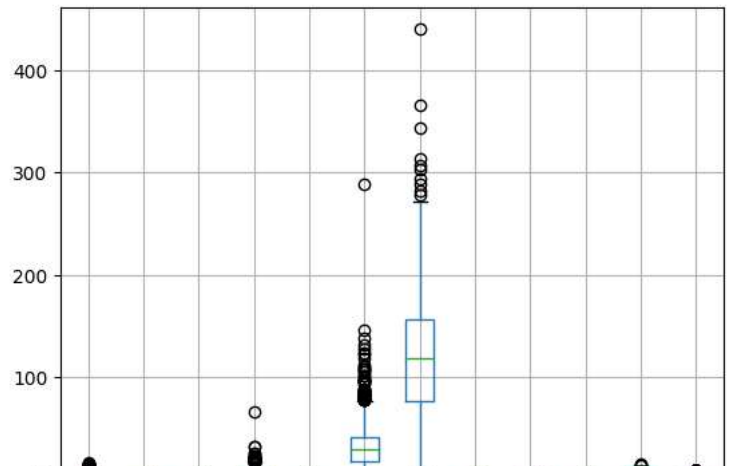
```
(6497, 13)
```

```
1 df.size
```

```
84461
```

```
1 df.boxplot()
```

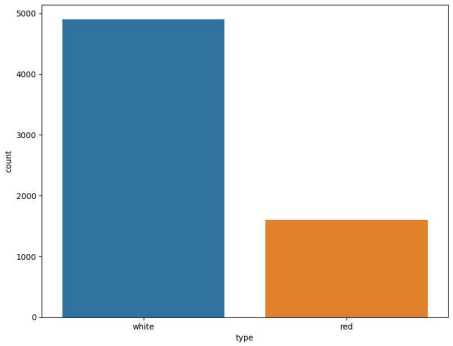
```
<Axes: >
```

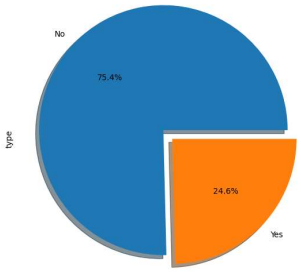



```
1 df["quality"].value_counts()
```

```
6    2836
5    2138
7    1079
4     216
8     193
3      30
9       5
Name: quality, dtype: int64
```

```
1 fig, ax = plt.subplots(1, 2, figsize=(20, 7))
2
3 sns.countplot(data=df, x="type", ax=ax[0])
4 df["type"].value_counts().plot.pie(explode=[0.1, 0],autopct="%1.1f%%", labels=["No", "Yes"], shadow=True, a
5
6 plt.show()
```






```
1 df.mean
```

```
<bound method NDFrame._add_numeric_operations.<locals>.mean of       type  fixed acidity  volatile acidity  citric acid  residual
sugar  \
0     white            7.0             0.270         0.36           20.7
1     white            6.3             0.300         0.34            1.6
2     white            8.1             0.280         0.40            6.9
3     white            7.2             0.230         0.32            8.5
4     white            7.2             0.230         0.32            8.5
...     ...            ...               ...          ...            ...
6492    red            6.2             0.600         0.08            2.0
6493    red            5.9             0.550         0.10            2.2
6494    red            6.3             0.510         0.13            2.3
6495    red            5.9             0.645         0.12            2.0
6496    red            6.0             0.310         0.47            3.6

      chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
```

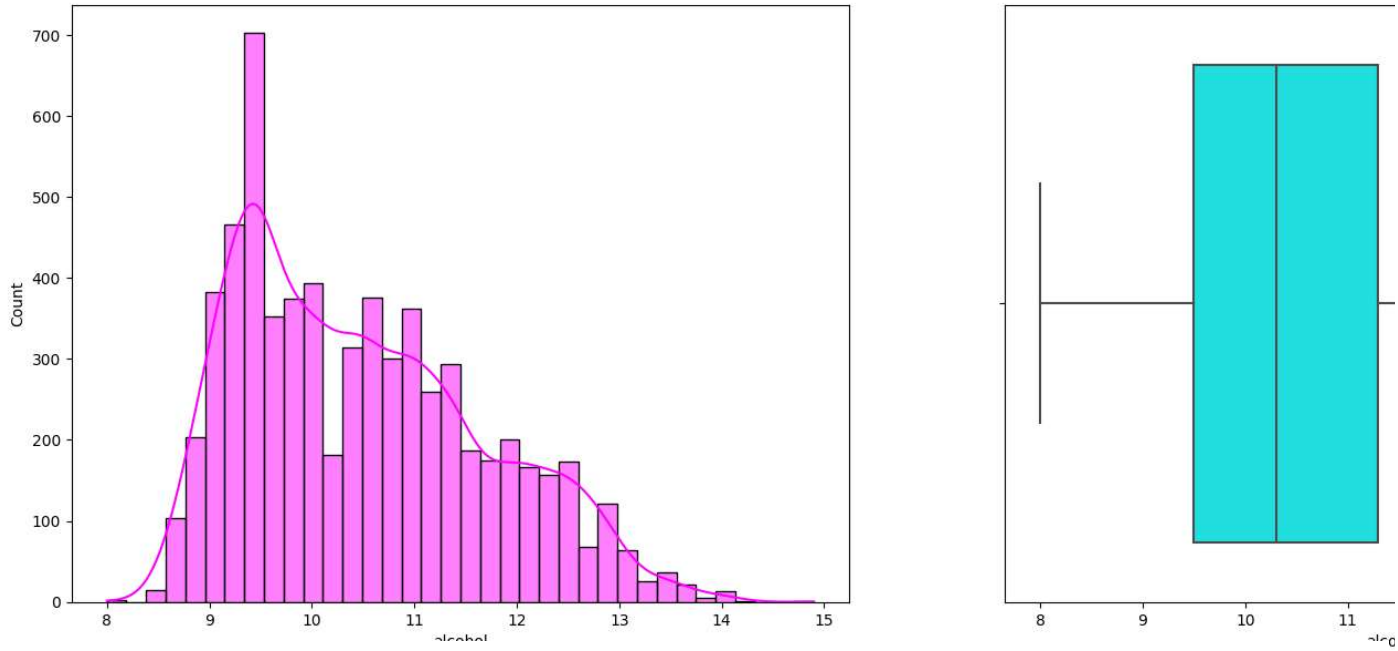
```
0         0.045                45.0                170.0  1.00100  3.00
1         0.049                14.0                132.0  0.99400  3.30
2         0.050                30.0                 97.0  0.99510  3.26
3         0.058                47.0                186.0  0.99560  3.19
4         0.058                47.0                186.0  0.99560  3.19
...         ...                 ...                  ...      ...   ...
6492      0.090                32.0                 44.0  0.99490  3.45
6493      0.062                39.0                 51.0  0.99512  3.52
6494      0.076                29.0                 40.0  0.99574  3.42
6495      0.075                32.0                 44.0  0.99547  3.57
6496      0.067                18.0                 42.0  0.99549  3.39

      sulphates  alcohol  quality
0          0.45      8.8        6
1          0.49      9.5        6
2          0.44     10.1        6
3          0.40      9.9        6
4          0.40      9.9        6
...         ...      ...      ...
6492       0.58     10.5        5
6493        NaN     11.2        6
6494       0.75     11.0        6
6495       0.71     10.2        5
6496       0.66     11.0        6

[6497 rows x 13 columns]>
```

```
1 df.median()
```

```
<ipython-input-17-6d467abf240d>:1: FutureWarning: The default value of numeric_only in DataFrame.median is deprecated. In a future vers
  df.median()
fixed acidity             7.00000
volatile acidity          0.29000
citric acid               0.31000
residual sugar            3.00000
chlorides                 0.04700
free sulfur dioxide      29.00000
total sulfur dioxide    118.00000
density                   0.99489
pH                        3.21000
sulphates                 0.51000
alcohol                  10.30000
quality                   6.00000
dtype: float64
```

```
1
2
3
4 sns.histplot(data=df, x="alcohol",kde=True,  ax=ax5[0],color = 'magenta')
5 sns.boxplot(data=df, x="alcohol", ax=ax5[1], color = 'aqua')
```

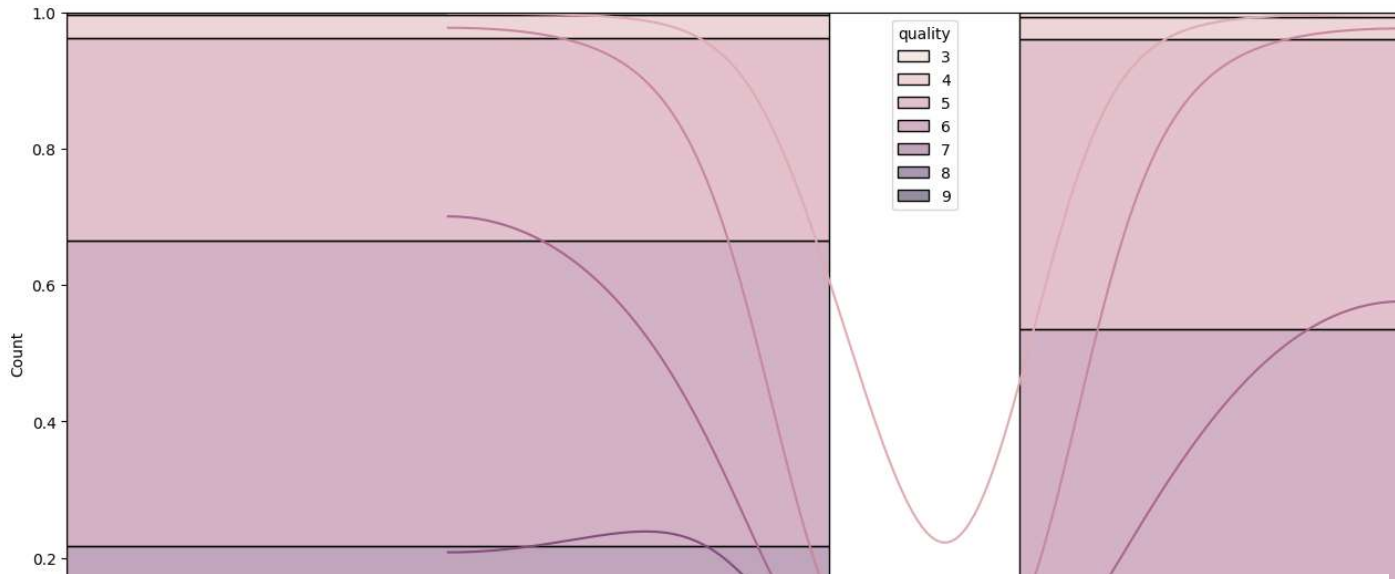
```
<Axes: xlabel='alcohol'>
```




```
1 print("Rows with alcohol value of 0: ", df[df["alcohol"] == 0].shape[0])
```

```
Rows with alcohol value of 0:  0
```

```
1 fig15, ax15 = plt.subplots(figsize=(20, 8))
2
3 sns.histplot(data=df, x="type", hue="quality", shrink=0.8, multiple="fill", kde=True, ax=ax15)
4 plt.show()
```



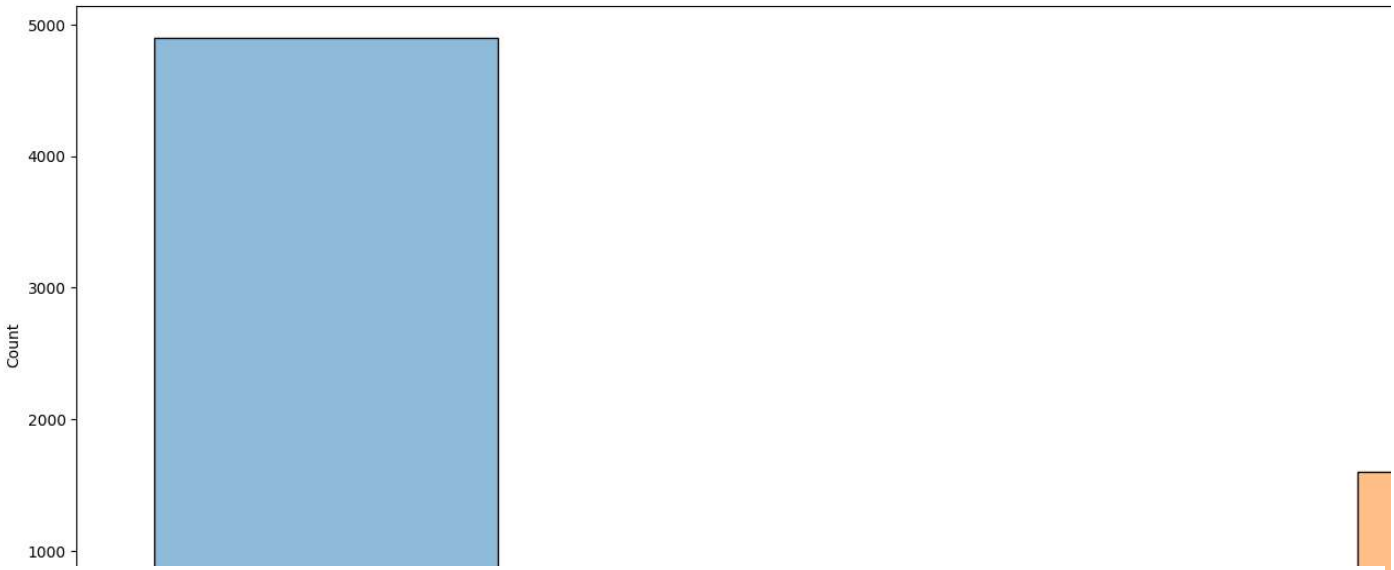

```
1 from matplotlib.ticker import MultipleLocator
2 fig16, ax16 = plt.subplots(figsize=(20, 8))
3
4 sns.histplot(data=df, x="type", hue="alcohol", shrink=0.8, multiple="dodge", kde=True, ax=ax16,)
5 plt.show()
```

```
df.isnull().sum()
```

```
type                    0
fixed acidity          10
volatile acidity        8
citric acid             3
residual sugar          2
chlorides               2
free sulfur dioxide     0
```

```
total sulfur dioxide     0
density                  0
pH                       9
sulphates                4
alcohol                  0
quality                  0
dtype: int64
```

```
fig16, ax16 = plt.subplots(figsize=(20, 8))

sns.histplot(data=df, x="type", hue="type", shrink=0.8, multiple="dodge", kde=True, ax=ax16,)
plt.show()

```




```
fig17, ax17 = plt.subplots(figsize=(20, 8))

sns.histplot(data=df, x="type", hue="type", shrink=0.8, multiple="fill", kde=True, ax=ax17)
plt.show()

```