# 20250529_01

May 29, 2025

[9]:
```
# Datasets come from UCI ML Repo
import pandas as pd
```

[7]:
```
# Load red wine
red_url = "https://archive.ics.uci.edu/ml/machine-learning-databases/
 ↪wine-quality/winequality-red.csv"
red = pd.read_csv(red_url, sep = ';')
red['type'] = 'red'
```

[6]:
```
# Load white wine
white_url = "https://archive.ics.uci.edu/ml/machine-learning-databases/
 ↪wine-quality/winequality-white.csv"
white = pd.read_csv(white_url, sep = ';')
white['type'] = 'white'
```

[8]:
```
# Combine datasets
wine = pd.concat([red, white], ignore_index = True)
```

[10]:
```
wine.head()
```

[10]:

| | 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 | | | |
| 4 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | | | |

```
3      9.8        6  red
4      9.4        5  red
```

[12]:
```
# No missing value
wine.info()
```

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

[21]:
```
wine[wine['type'] == 'red'].describe()
```

[21]:

| | fixed acidity | volatile acidity | citric acid | residual sugar | \ |
|---|---|---|---|---|---|
| count | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | |
| mean | 8.319637 | 0.527821 | 0.270976 | 2.538806 | |
| std | 1.741096 | 0.179060 | 0.194801 | 1.409928 | |
| min | 4.600000 | 0.120000 | 0.000000 | 0.900000 | |
| 25% | 7.100000 | 0.390000 | 0.090000 | 1.900000 | |
| 50% | 7.900000 | 0.520000 | 0.260000 | 2.200000 | |
| 75% | 9.200000 | 0.640000 | 0.420000 | 2.600000 | |
| max | 15.900000 | 1.580000 | 1.000000 | 15.500000 | |

| | chlorides | free sulfur dioxide | total sulfur dioxide | density | \ |
|---|---|---|---|---|---|
| count | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | |
| mean | 0.087467 | 15.874922 | 46.467792 | 0.996747 | |
| std | 0.047065 | 10.460157 | 32.895324 | 0.001887 | |
| min | 0.012000 | 1.000000 | 6.000000 | 0.990070 | |
| 25% | 0.070000 | 7.000000 | 22.000000 | 0.995600 | |
| 50% | 0.079000 | 14.000000 | 38.000000 | 0.996750 | |
| 75% | 0.090000 | 21.000000 | 62.000000 | 0.997835 | |
| max | 0.611000 | 72.000000 | 289.000000 | 1.003690 | |

```
                pH    sulphates      alcohol      quality
count  1599.000000  1599.000000  1599.000000  1599.000000
mean      3.311113     0.658149    10.422983     5.636023
std       0.154386     0.169507     1.065668     0.807569
min       2.740000     0.330000     8.400000     3.000000
25%       3.210000     0.550000     9.500000     5.000000
50%       3.310000     0.620000    10.200000     6.000000
75%       3.400000     0.730000    11.100000     6.000000
max       4.010000     2.000000    14.900000     8.000000
```

[22]:
```
wine[wine['type'] == 'white'].describe()
```

[22]:
```
       fixed acidity  volatile acidity  citric acid  residual sugar  \
count    4898.000000       4898.000000  4898.000000     4898.000000
mean        6.854788          0.278241     0.334192        6.391415
std         0.843868          0.100795     0.121020        5.072058
min         3.800000          0.080000     0.000000        0.600000
25%         6.300000          0.210000     0.270000        1.700000
50%         6.800000          0.260000     0.320000        5.200000
75%         7.300000          0.320000     0.390000        9.900000
max        14.200000          1.100000     1.660000       65.800000

         chlorides  free sulfur dioxide  total sulfur dioxide      density  \
count  4898.000000          4898.000000           4898.000000  4898.000000
mean      0.045772            35.308085            138.360657     0.994027
std       0.021848            17.007137             42.498065     0.002991
min       0.009000             2.000000              9.000000     0.987110
25%       0.036000            23.000000            108.000000     0.991723
50%       0.043000            34.000000            134.000000     0.993740
75%       0.050000            46.000000            167.000000     0.996100
max       0.346000           289.000000            440.000000     1.038980

                pH    sulphates      alcohol      quality
count  4898.000000  4898.000000  4898.000000  4898.000000
mean      3.188267     0.489847    10.514267     5.877909
std       0.151001     0.114126     1.230621     0.885639
min       2.720000     0.220000     8.000000     3.000000
25%       3.090000     0.410000     9.500000     5.000000
50%       3.180000     0.470000    10.400000     6.000000
75%       3.280000     0.550000    11.400000     6.000000
max       3.820000     1.080000    14.200000     9.000000
```

[14]:
```
# Doing T-test on pH level between red and white wine
# H  : Mean pH is equal for red and white wines
# H  : Mean pH is different
from scipy.stats import ttest_ind
```

[29]:
```
# Extract pH level
red_pH = wine[wine['type'] == 'red']['pH']
white_pH = wine[wine['type'] == 'white']['pH']

# Perform Welch's t-test, since they have different sample size.
# equal_var can be True(Student) or False(Welch)
t_stat, p_val = ttest_ind(red_pH, white_pH, equal_var = False)

print(f"t-statistic: {t_stat:.4f}")
print(f"p-value: {p_val:.4e}")
```

```
t-statistic: 27.7755
p-value: 2.3423e-149
```

So there **IS** difference between the **pH level** of these two kinds of wine, but not by much. Since **Red_pH_mean** = 3.311113, and **White_pH_mean** = 3.188267

[30]:
```
# Now we do T-test on wine quality score between red and white wine
# H  : Mean quality score is the same for red and white wine
# H  : Mean quality score is different

# Extract quality scores
red_quality = wine[wine['type'] == 'red']['quality']
white_quality = wine[wine['type'] == 'white']['quality']

# Welch's t-test again
t_stat, p_val = ttest_ind(red_quality, white_quality, equal_var = False)

print(f"t-statistic: {t_stat:.4f}")
print(f"p-value: {p_val:.4e}")
```

```
t-statistic: -10.1494
p-value: 8.1683e-24
```

Since P-value $< 0.05$, we say that there is a **difference** in average quality score between red and white wine, which **white wine** is slightly better.