DSBDL Assignment 03 - Descriptive Statistics: Measures of Central Tendency and Variability

Part 1

Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variable. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable.

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
ds = pd.read_csv('/content/drive/My Drive/DSBDL/Assignment3/winequality-white.csv', sep=';')
ds
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.00100	3.00
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.99400	3.30
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.99510	3.26
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.99560	3.19
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.99560	3.19
...	...	...	...	...	...	...	...	...	...
4893	6.2	0.21	0.29	1.6	0.039	24.0	92.0	0.99114	3.27
4894	6.6	0.32	0.36	8.0	0.047	57.0	168.0	0.99490	3.15
4895	6.5	0.24	0.19	1.2	0.041	30.0	111.0	0.99254	2.99
4896	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34
4897	6.0	0.21	0.38	0.8	0.020	22.0	98.0	0.98941	3.26

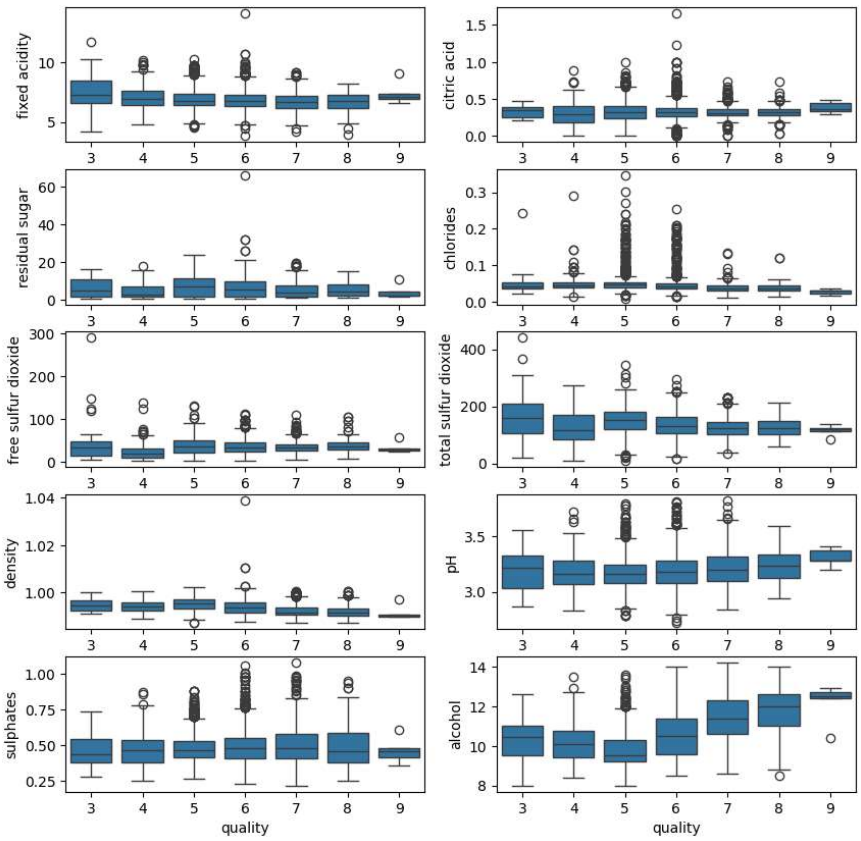
Next steps: Generate code with ds | View recommended plots

```
ds.drop( [ "quality" ] , axis=1 ).describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	48
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	1
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	1
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	1
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	1

```
numeric_features = [ "fixed acidity" ,
"citric acid" ,
"residual sugar" ,
"chlorides" ,
"free sulfur dioxide" ,
"total sulfur dioxide" ,
"density" ,
"pH" ,
"sulphates" ,
"alcohol" ]

fig, axes = plt.subplots(5 , 2 , figsize=( 10 , 10 ))
axes = axes.flatten()
for i , feature in enumerate( numeric_features ):
    sns.boxplot( data=ds , y=feature, x="quality" , ax=axes[i] )
```




```
def group_stats( feature_name ):
    labels = ds[ "quality" ].unique().tolist()
    for label in labels:
        print( f"Label: {label}")
        print( ds[ ds[ "quality" ] == label ][ feature_name ].describe() , end="\n\n" )

for feature in numeric_features:
    group_stats( feature )
```

```
Label: 8
count    175.000000
mean      11.636000
std        1.280138
min        8.500000
25%       11.000000
50%       12.000000
75%       12.600000
max       14.000000
Name: alcohol, dtype: float64

Label: 4
count    163.000000
mean      10.152454
std        1.003217
min        8.400000
25%        9.400000
50%       10.100000
75%       10.750000
max       13.500000
Name: alcohol, dtype: float64

Label: 3
count    20.000000
mean     10.345000
std       1.224089
min       8.000000
25%       9.550000
50%      10.450000
75%      11.000000
max      12.600000
Name: alcohol, dtype: float64

Label: 9
count     5.00000
mean     12.18000
std       1.01341
min      10.40000
25%      12.40000
50%      12.50000
75%      12.70000
max      12.90000
Name: alcohol, dtype: float64
```