WINE QUALITY

[[reference link_dataset]]

Establishing python packages

Show code

Dataset Extraction from CSV File

```
# @title Dataset Extraction from CSV File
# extraction of data set from csv file; also removing row indexes
raw = pd.read_csv('/winequalityN.csv')
raw.head()
```

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

Next steps: View recommended plots

TRANSFORMING AND MANIPULATING CSV DATAFRAME

Cleaning and Transforming Frames

Show code

```
DATAFRAME(transformed) OVERVIEW
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 14 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
 13  quality_label         6497 non-null   category
dtypes: category(1), float64(11), int64(1), object(1)
memory usage: 666.5+ KB
```

Dataset Overview

Show code

	type	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	fre
0	white	7.0	0.27	0.36	20.7	0.045	
1	white	6.3	0.30	0.34	1.6	0.049	
2	white	8.1	0.28	0.40	6.9	0.050	
3	white	7.2	0.23	0.32	8.5	0.058	
4	white	7.2	0.23	0.32	8.5	0.058	

Next steps: View recommended plots

Exploring DataFrame: identifying identicals according to columns 'type'

Show code

```
  Wine_Type  Count
0     white   4898
1       red   1599

Sum of Counted Data Entries for white and red wines: 6497
Sum of Data Entries (rows) from Data Set: 6497
```

Type Column Summary

Show code

```
Number of data set (in rows) according to wine types:
White Wine = 4898
```

```
Red Wine = 1599
```

DataFrame: Type White Wine

Show code

	type	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides
0	white	7.0	0.27	0.36	20.7	0.045
1	white	6.3	0.30	0.34	1.6	0.049
2	white	8.1	0.28	0.40	6.9	0.050
3	white	7.2	0.23	0.32	8.5	0.058
4	white	7.2	0.23	0.32	8.5	0.058
...	...	...	...	...	...	...
4893	white	6.2	0.21	0.29	1.6	0.039
4894	white	6.6	0.32	0.36	8.0	0.047
4895	white	6.5	NaN	0.19	1.2	0.041
4896	white	5.5	0.29	0.30	1.1	0.022
4897	white	6.0	0.21	0.38	0.8	0.020

4898 rows × 14 columns

Next steps: View recommended plots

Identifyting 'NaN' Values for Type: White Wine

Show code

```
Count of NaN values in df_type_white:
type                  0
fixed_acidity         8
volatile_acidity      7
citric_acid           2
residual_sugar        2
chlorides             2
free_sulfur_dioxide   0
total_sulfur_dioxide  0
density               0
pH                    7
sulphates             2
alcohol               0
quality               0
```

```
quality_label          0
dtype: int64
```

Cleaning Dataframe: Type White Wine

Show code

```
DataFrame after removing rows with NaN values:
      alcohol  chlorides  citric_acid  density  fixed_acidity  \
0         8.8      0.045         0.36  1.00100            7.0
1         9.5      0.049         0.34  0.99400            6.3
2        10.1      0.050         0.40  0.99510            8.1
3         9.9      0.058         0.32  0.99560            7.2
4         9.9      0.058         0.32  0.99560            7.2
...       ...        ...          ...      ...            ...
4865     10.6      0.038         0.32  0.99074            5.7
4866     11.2      0.039         0.29  0.99114            6.2
4867      9.6      0.047         0.36  0.99490            6.6
4868     12.8      0.022         0.30  0.98869            5.5
4869     11.8      0.020         0.38  0.98941            6.0

      free_sulfur_dioxide    pH  quality quality_label  residual_sugar  \
0                    45.0  3.00        6        medium            20.7
1                    14.0  3.30        6        medium             1.6
2                    30.0  3.26        6        medium             6.9
3                    47.0  3.19        6        medium             8.5
4                    47.0  3.19        6        medium             8.5
...                   ...   ...      ...           ...             ...
4865                 38.0  3.24        6        medium             0.9
4866                 24.0  3.27        6        medium             1.6
4867                 57.0  3.15        5           low             8.0
4868                 20.0  3.34        7        medium             1.1
4869                 22.0  3.26        6        medium             0.8

      sulphates  total_sulfur_dioxide   type  volatile_acidity
0          0.45                 170.0  white              0.27
1          0.49                 132.0  white              0.30
2          0.44                  97.0  white              0.28
3          0.40                 186.0  white              0.23
4          0.40                 186.0  white              0.23
...         ...                   ...    ...               ...
4865       0.46                 121.0  white              0.21
4866       0.50                  92.0  white              0.21
4867       0.46                 168.0  white              0.32
4868       0.38                 110.0  white              0.29
4869       0.32                  98.0  white              0.21

[4870 rows x 14 columns]
```

EXPLORING DATASET: for type = white wine

Statistical Overview of Dataset: Type White Wine

Show code

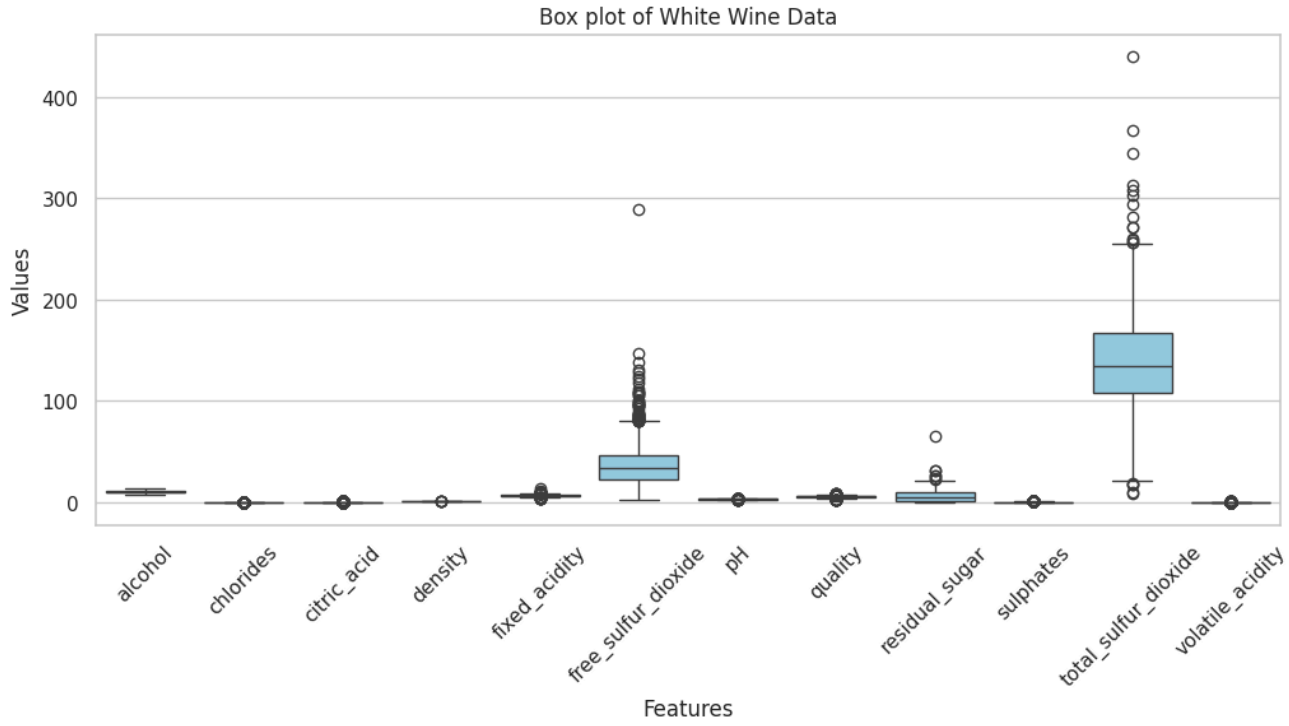
	alcohol	chlorides	citric_acid	density	fixed_acidity	free_sulfur_dioxide
count	4870.00	4870.00	4870.00	4870.00	4870.00	4870.00
mean	10.52	0.05	0.33	0.99	6.86	35.32
std	1.23	0.02	0.12	0.00	0.84	17.01
min	8.00	0.01	0.00	0.99	3.80	2.00
25%	9.50	0.04	0.27	0.99	6.30	23.00
50%	10.40	0.04	0.32	0.99	6.80	34.00
75%	11.40	0.05	0.39	1.00	7.30	46.00
max	14.20	0.35	1.66	1.04	14.20	289.00

NOTES:

- having an erratic standard deviation across different attributes, it can be inferred that the row dataset MAY NOT represent only 1 wine formulation. Hence, target: identify which row dataset are quite similar to another and different from one another
- for Exploratory Analysis: given the highest standard deviation value, investigate attribute 'total_sulfur_dioxide'

Box Plot Overview: White Wine dataset across 12 attributes (regardless of quality type)

Show code



Box Plot above visualizes how values greatly deviates from the other in attributes (1) total_sulfur_dioxide, (2) free_sulfur_dioxide

DataFrame: [high, medium, low] Qualities, [white] Type Wines

Show code

To identify which row dataset are quite similar to another and different from one another, [WHITE] Type wines were sorted according to qualities [high], [medium]], [low]

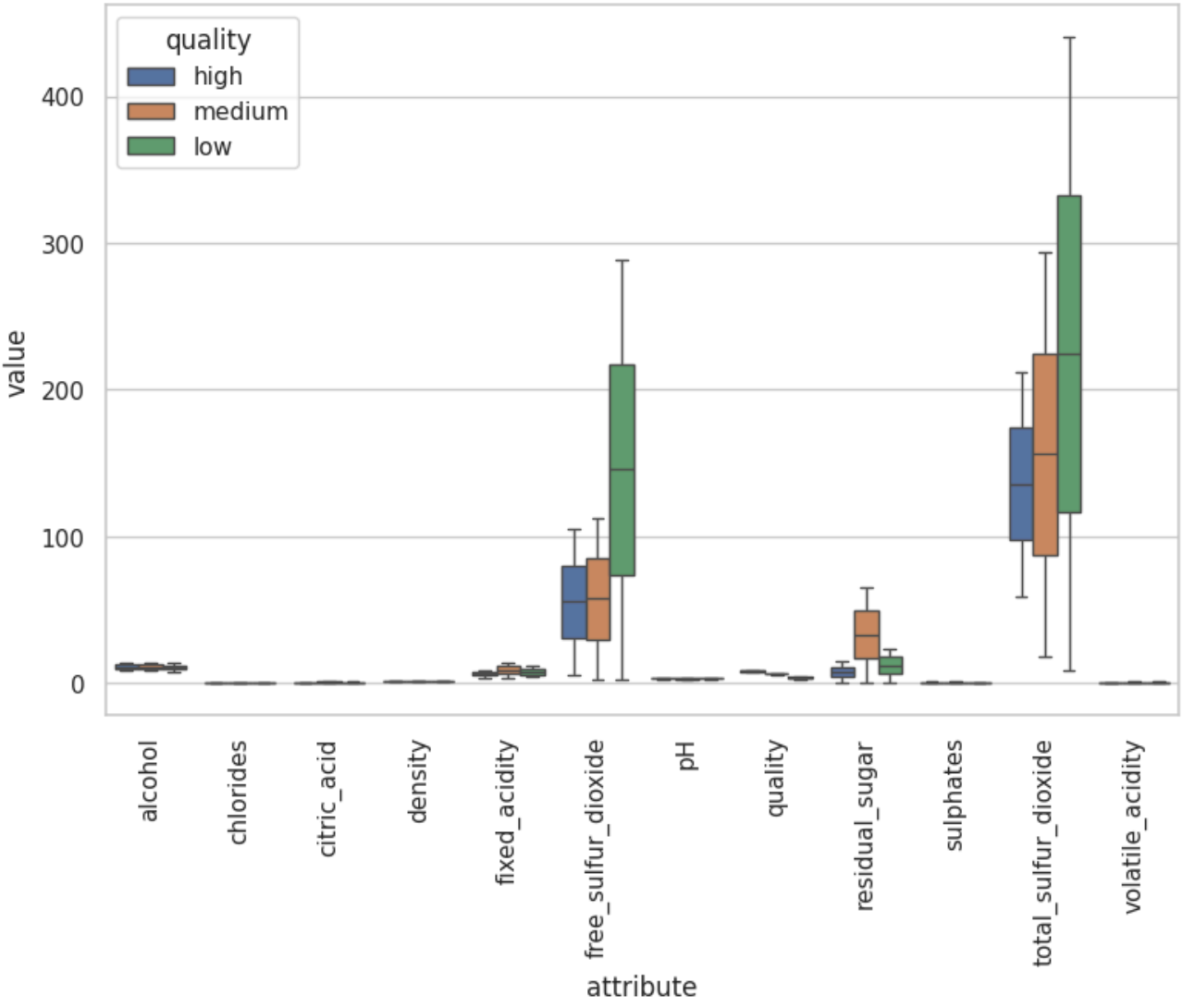
Table of Max and Min Values per attribute for the Multiple Box Plot

Show code

Box Plot according to Quality Types of [White] Type Wines

Show code

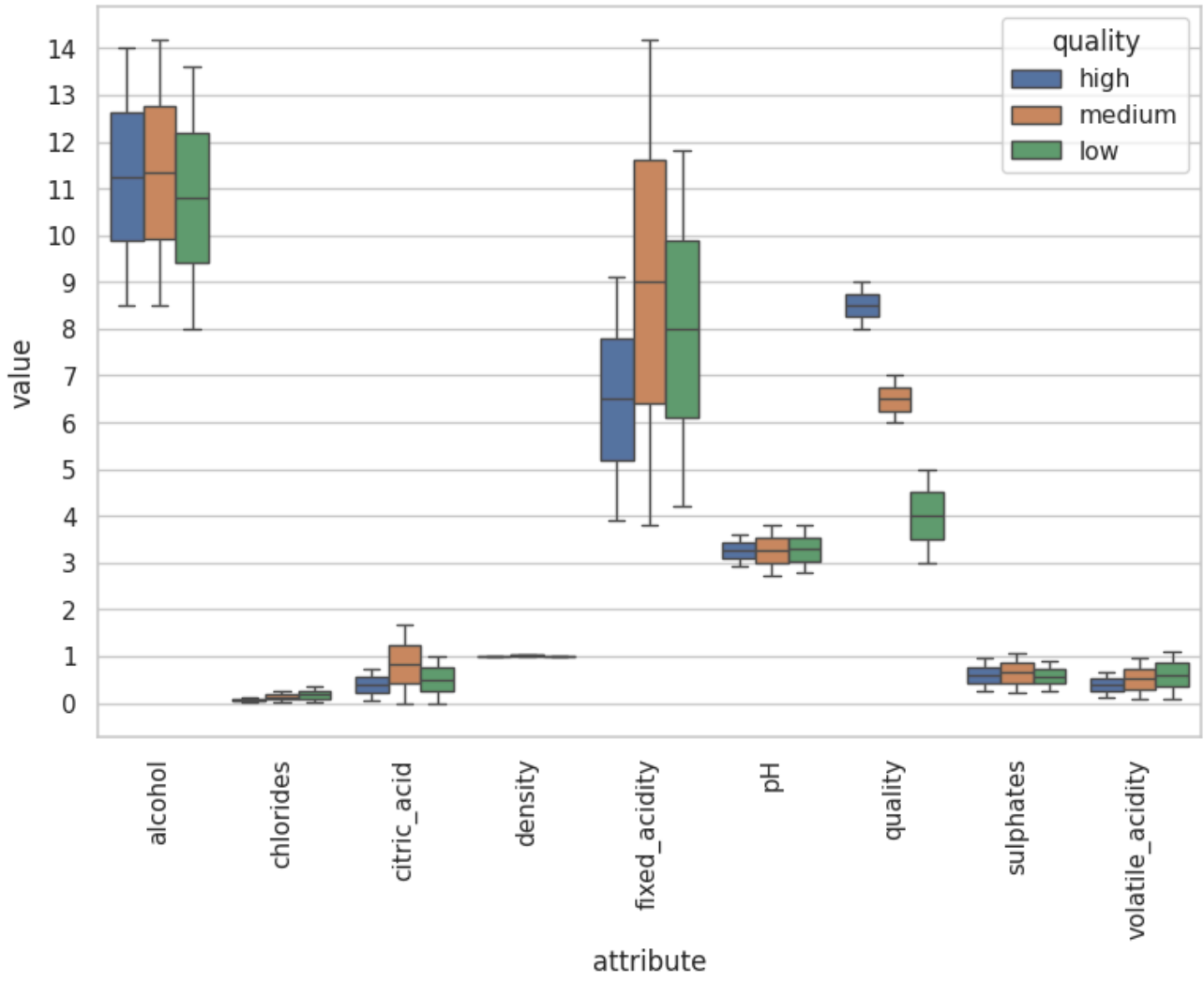
```
<ipython-input-70-c53da1a9a7a2>:29: UserWarning: FixedFormatter should only be used t
  ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
```



Zooming into non-extreme attributes: all except sulfur_dioxides and 'residual_sugar'

Show code

```
<ipython-input-70-c53da1a9a7a2>:29: UserWarning: FixedFormatter should only be used t
  ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
```



EXLORATORY ANALYSIS: 'total_sulfur_dioxide'

Overview of Working Dataframe: for (type) White Wine; (attribute) Total Sulfur Dioxide

Show code

```
Median Value: 134.0
count    4870.00
mean      138.34
std        42.49
min         9.00
25%       108.00
50%       134.00
```

```
75%       167.00
max       440.00
Name: total_sulfur_dioxide, dtype: float64
```

Quartiles: Generating Table of Values according to desired inputted quartile range

Show code

SUMMARY

- Q1 = 108.0 [1229 frequency]
- Q2 = 134.0 [1227 frequency]
- Q3 = 167.0 [1203 frequency]
- Q4 = 440.0 [1211 frequency]

Total values (frequency): 4870

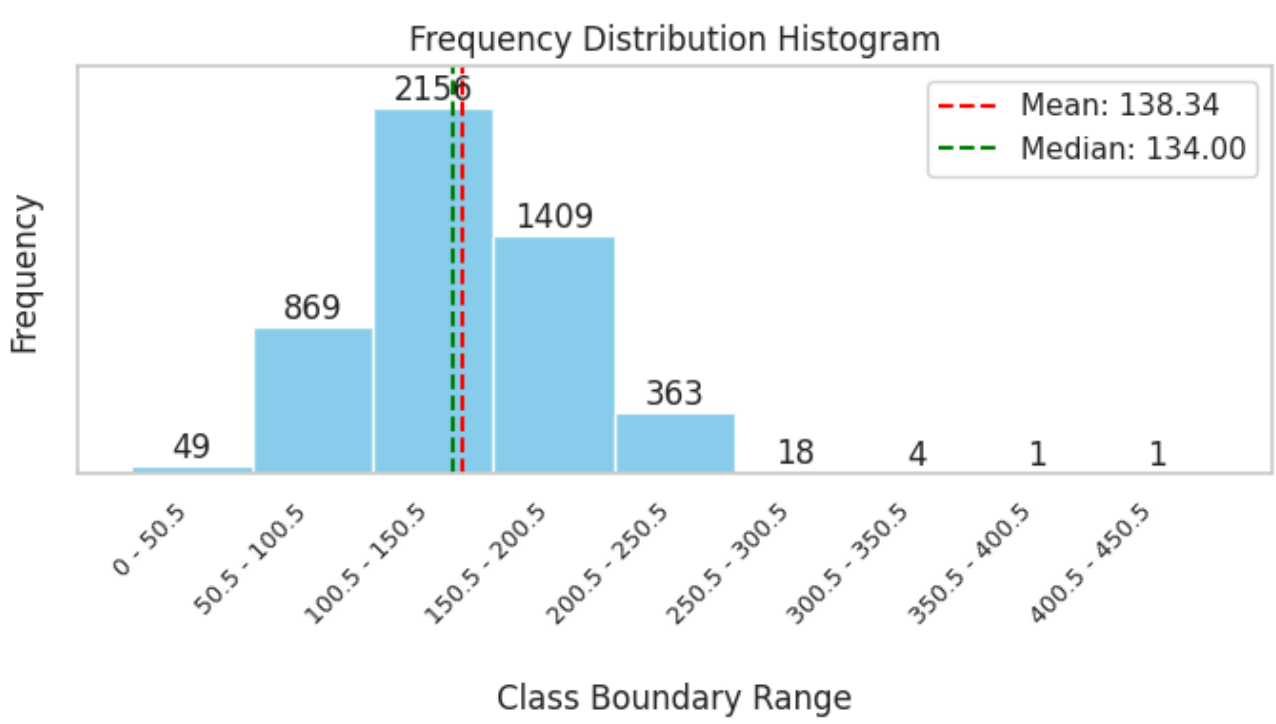
Frequency Distribution Table

Show code

	Lower_Class_Boundary	Upper_Class_Boundary	Frequency
0 - 50.5	0.0	50.5	49
50.5 - 100.5	50.5	100.5	869
100.5 - 150.5	100.5	150.5	2156
150.5 - 200.5	150.5	200.5	1409
200.5 - 250.5	200.5	250.5	363
250.5 - 300.5	250.5	300.5	18
300.5 - 350.5	300.5	350.5	4
350.5 - 400.5	350.5	400.5	1
400.5 - 450.5	400.5	450.5	1

Frequency Distribution: Histogram

Show code



// REMOVAL OF OUTLIERS AND FOLOW-UP ANALYSIS

Ver 1 Cleaning: Removing Outliers

Show code

```
From 4870 rows to 4864
Updated Mean Value: 138.08
Updated Median Value: 134.0
3684      9.0
3875     10.0
3069     18.0
3068     18.0
721      19.0
        ...
375     260.0
219     272.0
3024    272.0
2354    282.0
3126    294.0
Name: total_sulfur_dioxide, Length: 4864, dtype: float64
```

Frequency Distribution Table (exc. 6 outliers; increments by 30s) [ver 1]

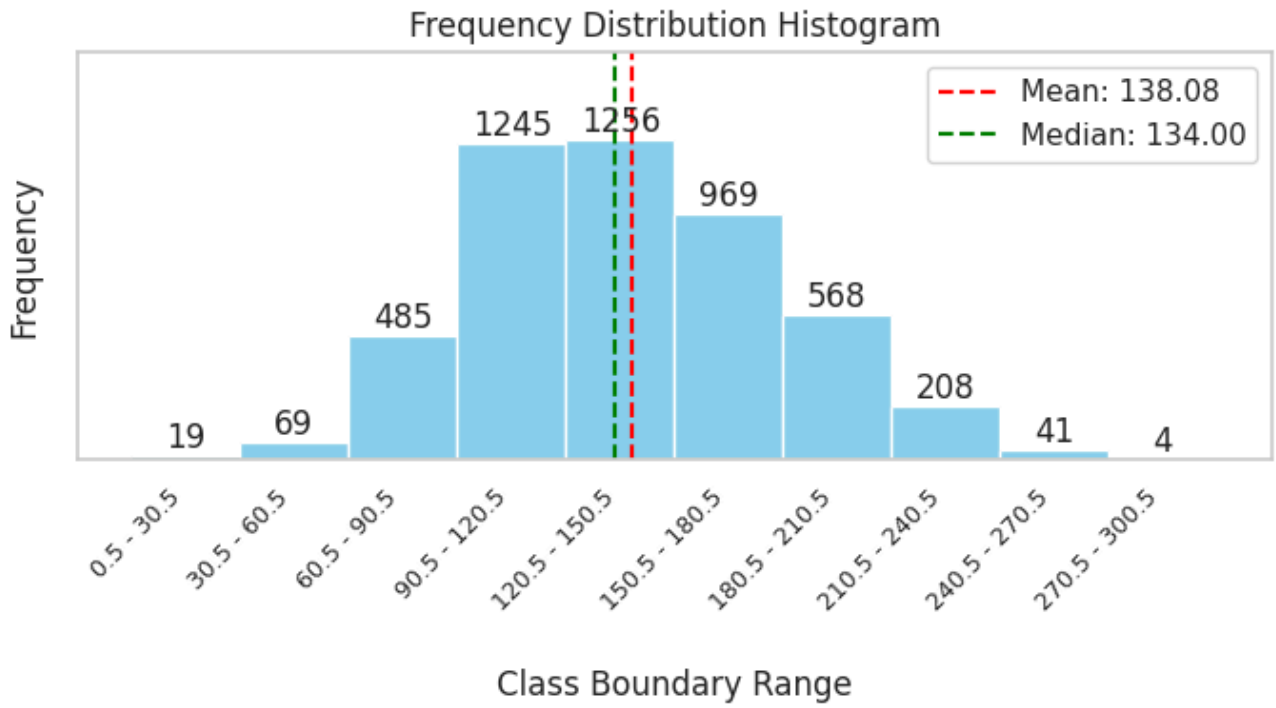
Show code

	Lower_Class_Boundary	Upper_Class_Boundary	Frequency
0.5 - 30.5	0.5	30.5	19
30.5 - 60.5	30.5	60.5	69

```
60.5 - 90.5      60.5      90.5      485
90.5 - 120.5     90.5      120.5     1245
120.5 - 150.5    120.5     150.5     1256
150.5 - 180.5    150.5     180.5     969
180.5 - 210.5    180.5     210.5     568
210.5 - 240.5    210.5     240.5     208
240.5 - 270.5    240.5     270.5     41
270.5 - 300.5    270.5     300.5     4
```

Frequency Distribution: Histogram (exc. 6 outliers; increments by 30s) [ver 1]

Show code



Ver 2 Cleaning: Removing Outliers

Show code

```
From 4870 rows, to 4864, to 4860
Updated Mean Value: 137.97
Updated Median Value: 134.0
3684      9.0
3875     10.0
3069     18.0
3068     18.0
721      19.0
        ...
106     255.0
1916    256.0
1918    256.0
4488    259.0
```

```
375     260.0
Name: total_sulfur_dioxide, Length: 4860, dtype: float64
```

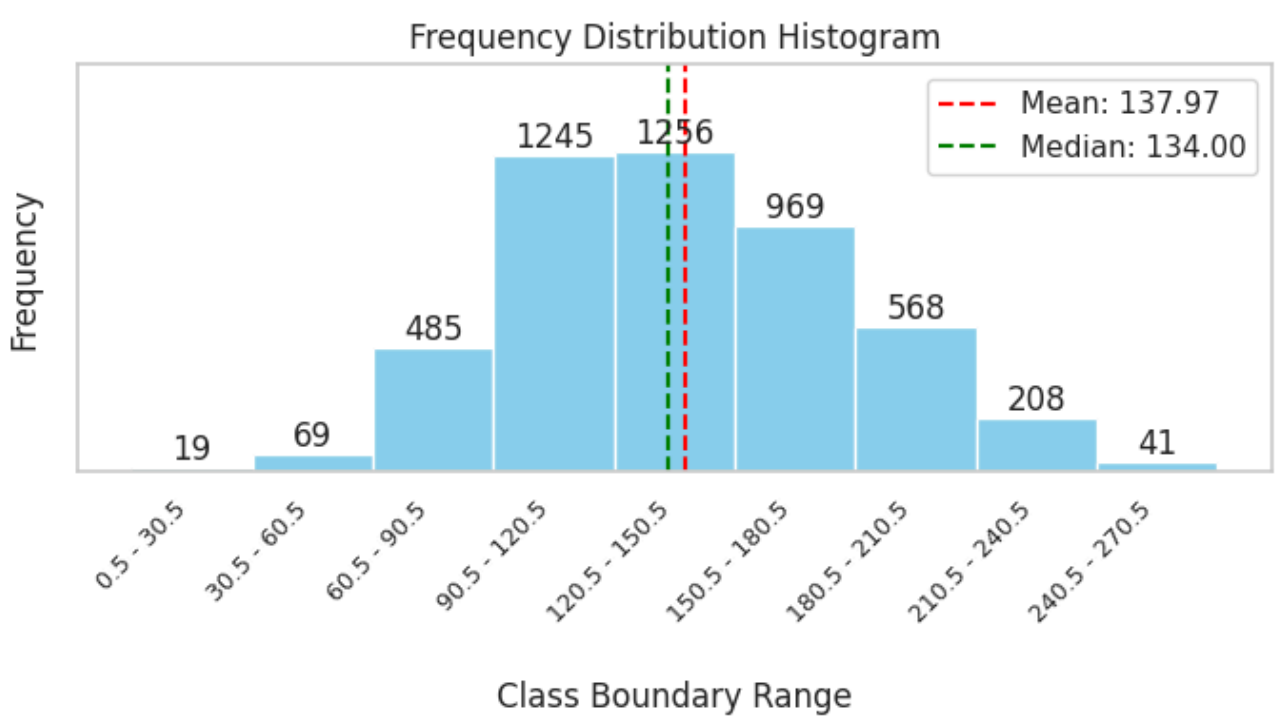
Frequency Distribution Table (exc. 6 outliers; increments by 30s) [ver 2]

Show code

	Lower_Class_Boundary	Upper_Class_Boundary	Frequency
0.5 - 30.5	0.5	30.5	19
30.5 - 60.5	30.5	60.5	69
60.5 - 90.5	60.5	90.5	485
90.5 - 120.5	90.5	120.5	1245
120.5 - 150.5	120.5	150.5	1256
150.5 - 180.5	150.5	180.5	969
180.5 - 210.5	180.5	210.5	568
210.5 - 240.5	210.5	240.5	208
240.5 - 270.5	240.5	270.5	41

Frequency Distribution: Histogram (exc. 6 outliers; increments by 30s) [ver 2]

Show code



Indexes of deemed 10 outliers in white wine (type), 'total_sulfur_dioxide'(attribute)

Show code

```
Note: index numbers of rows containing outliers with respect to total_sulfur_dioxide
[219, 314, 1393, 1907, 2103, 2354, 2630, 3024, 3126, 4719]
```

EXPLORATORY ANALYSIS: 'quality_label'

Overview of Working Dataframe: for (type) White Wine; (attribute) Quality Label

Show code

```
Low Quality Frequency: 1630
Medium Quality Frequency: 3061
High Quality Frequency: 179
4870
```

To enumerate column names as a tabled list

Show code

	0
0	alcohol
1	chlorides
2	citric_acid