Name: Tahmim Hassan (tahmim93.hassan@gmail.com)

Summary Notes:

If Google Colab is used to run the files, it is advisable to press ‘Run all’ from the above, from the ‘Runtime’ tab, so the pointer can run the data sequentially.

Section 1: Data Cleaning:

There were two datasets in the file on the mentioned link: winequality-red and winequality-white. I opened up the files in Google Sheets and downloaded each of the two CSV files from Sheets respectively.

Data Cleansing and **Visualization** is carried out for both of the datasets respectively.

- There were no null (NaN/ NA) values in any of the datasets, in any of the columns. So imputing or interpolation was not needed.

- **Duplicates:**

There were quite some duplicate entries in both the datasets (all features had to be the same for two or more entries to be considered duplicates).

- There were **240** duplicate entries in the winequality-red dataset.
- There were **937** duplicate entries in the winequality-white dataset.

Duplicate entries were dropped from the table.

- The **‘quality’** feature was of **int** type initially, and it was changed to **float** type to maintain consistency and data integrity throughout all values in the dataset.

- There were also no categorical variables in the data. So no encoding was needed. There were also no date formats in the dataset.

Outliers are extreme values which deviate from the normal (average) pattern of values and outliers were detected here using the **Interquartile** range. Values which deviate significantly from the normal can be imputed with mean, median etc, or can be removed to improve the accuracy and consistency of data. In this case, outlier values were removed (as asked in the question).
Outlier values are replaced with mean/median values if dropping them reduces a lot of data from the dataset.

Before outliers were removed, the dataset had 1599 rows and 12 columns. After removal of outlier values, there were 1124 rows and 12 columns in the winequality-red file. The winequality-white file had 4898 rows and 12 columns previously, and after removal of outliers it had 3961 rows and 12 columns.

Outliers of two variables ‘total sulfur dioxide’ and ‘residual sugar’ are also visualized using **box plot diagrams**.

Section 2: Data Visualization:

Number 1; Barchart: A new variable ‘quality_categorical’, was made based on the variable ‘quality’.
Quality value of 0-3 was taken as ‘poor', 4-6 was taken as ‘average’ and a quality value greater than 6 (from 7 to 10) meant it was ‘good’.
A barchart was plotted accordingly.

Number 2; Line Plot: The trends of numerical variables ‘alcohol’ and ‘pH’ is visualized using line plots.

Number 3; Scatter Plot: A scatter plot of two numerical variables, ‘pH’, and ‘total sulfur dioxide’ was plotted to demonstrate their relationships. Firstly a scatter plot including outlier values is plotted. Then a scatter plot with the outlier values removed is plotted, which shows a very uniform and balanced distribution of data points.

Number 4; Pie Chart: A Pie Chart is drawn to represent the proportion of different categories in a dataset. Only the unique values of each variable is considered. Also the variable ‘quality’ is not considered in the making of pie chart as it is a mostly qualitative variable (Quality of the wine), and also its proportion size is quite small.

Number 5; Heatmap: A heatmap was made using Seaborn to represent the correlation between different numerical variables. A variable will have 100% correlation with itself as demonstrated by the diagonal blocks of values of 1 in the middle. E.g. Variable ‘fixed acidity’ will have complete correlation with ‘fixed acidity’ (itself) so its correlation value is 1.

Furthermore, a profile report could be generated from pandas by using Pandas Profiling. Pandas can generate a profile report file by itself which demonstrates statistical details about the data. Sometimes Google Colab can run out of memory and give a MemoryError for this block of code, so I commented that entire block of the code (in the last).