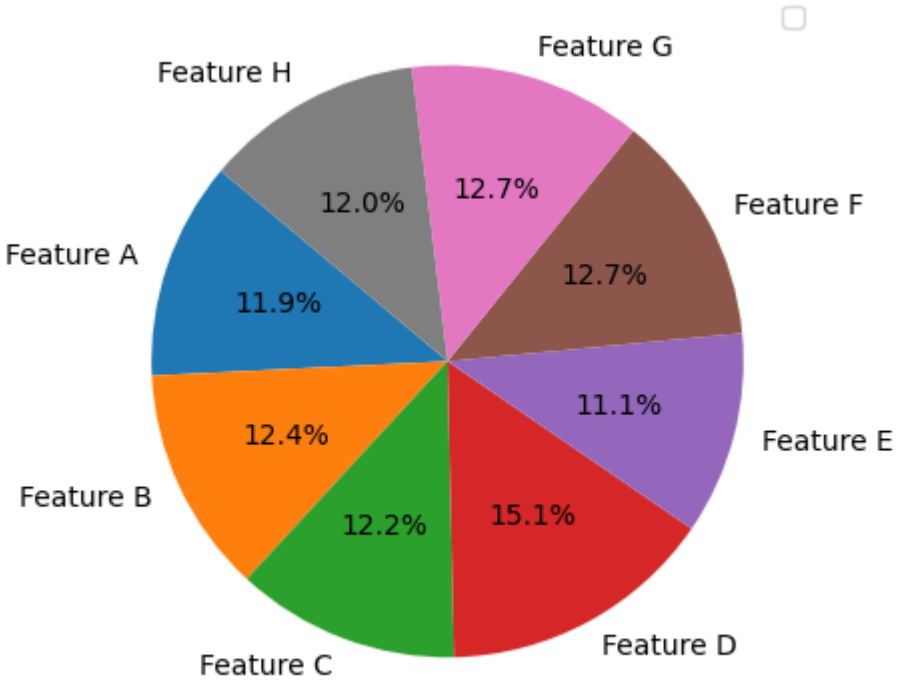
Q1:

Preprocessing Tasks

1. Cleaning
    a. There are some rows which do not have values for some columns, so I choose to replace those missing values with the mean value of the column.
    b. There is whitespace in some of the ordinal columns, so I choose to remove the whitespace
    c. Some ordinal values have uppercase and lowercase letters, so I choose to normalize them all to uppercase
2. Transformation
    a. Removed outliers using z-score and a threshold of 2
    b. Discretized features C and D
    c. Normalized data using min-max

Q2:

I used the pandas, numpy, and scipy python libraries to more easily process the data using the processes outlined in Q1.



Q3:

```
Pearson correlation between features and target variable:
acidity                -0.113473
volatile acidity       -0.194702
citric acid            -0.008786
residual sugar         -0.097684
chlorides              -0.209247
total sulfur dioxide   -0.174556
density                -0.306731
pH                      0.099841
sulphates               0.053166
```

```
Ranked features based on correlation with target variable:
alcohol                0.434807
density                0.306731
chlorides              0.209247
volatile acidity       0.194702
total sulfur dioxide   0.174556
acidity                0.113473
pH                     0.099841
residual sugar         0.097684
sulphates              0.053166
citric acid            0.008786
```

```
Correlation matrix for all features:
                      acidity  volatile acidity  citric acid  residual sugar  chlorides  total sulfur dioxide   density        pH  sulphates   alcohol   quality
acidity              1.000000         -0.022857     0.289218        0.088892   0.023042             0.091045  0.265265 -0.425586  -0.017251 -0.120921 -0.113473
volatile acidity    -0.022857          1.000000    -0.149758        0.064357   0.070539             0.089033  0.027120 -0.032192  -0.035703  0.067731 -0.194702
citric acid          0.289218         -0.149758     1.000000        0.094243   0.114345             0.121485  0.149537 -0.163472   0.062541 -0.075650 -0.008786
residual sugar       0.088892          0.064357     0.094243        1.000000   0.088753             0.401528  0.838960 -0.194280  -0.026310 -0.450440 -0.097684
chlorides            0.023042          0.070539     0.114345        0.088753   1.000000             0.198928  0.257239 -0.090498   0.016869 -0.360132 -0.209247
total sulfur dioxide 0.091045          0.089033     0.121485        0.401528   0.198928             1.000000  0.529885  0.002321   0.134944 -0.448627 -0.174556
density              0.265265          0.027120     0.149537        0.838960   0.257239             0.529885  1.000000 -0.093651   0.074659 -0.780022 -0.306731
pH                  -0.425586         -0.032192    -0.163472       -0.194280  -0.090498             0.002321 -0.093651  1.000000   0.155682  0.121320  0.099841
sulphates           -0.017251         -0.035703     0.062541       -0.026310   0.016869             0.134944  0.074659  0.155682   1.000000 -0.017244  0.053166
alcohol             -0.120921          0.067731    -0.075650       -0.450440  -0.360132            -0.448627 -0.780022  0.121320  -0.017244  1.000000  0.434807
quality             -0.113473         -0.194702    -0.008786       -0.097684  -0.209247            -0.174556 -0.306731  0.099841   0.053166  0.434807  1.000000
```

```
Redundant features (correlation >= 0.65):
('density', 'residual sugar')
('alcohol', 'density')
```