Question 1: Remove the rows with missing labels ('class') and rows with more than 7 missing features. Report the remaining number of rows. (1 mark)

The original number of rows is 178 and remaining number of rows is 154.

Question 2: Remove features with > 50% of missing values. For other features with missing values fill them with the mean of the corresponding features. Report the removed features (if any) and standard deviation of features with missing values after filling. (2 marks)

The removed feature is "Ash" and the following is the standard deviation of features with mi ssing value after filling

```
class                   0.766522
Alcohol                 3.804067
Malic acid              1.116005
Alcalinity of ash       3.456794
Magnesium              14.440377
Total phenols           0.617237
Flavanoids              0.873573
Nonflavanoid phenols    0.127083
Proanthocyanins         0.587671
Color intensity         2.325204
Hue                     0.229412
OD280/OD315             0.723261
Proline               303.033368
```

Question 3: Detect and remove rows with any outliers/incorrect values in features 'alcohol' and 'proline' (if any). Clearly state the basis of your removal. (1 mark)

In this problem, we remove the outliers based on the difference between value and feature mean larger than 3 times of feature's standard deviation

Question 4: Train Decision Tree model on train data for criterions = ('gini', 'entropy') and report the accuracies on the validation data. Select the best criterion and report the accuracy on the test data. (1 mark)

Answer:
gini: Validation accuracy = 0.974358974359
entropy: Validation accuracy = 0.948717948718

```
Test accuracy = 0.769230769231
```

Our results showed that the accuracy on the validataion data is 95% by using Decision Tree model on train data for entropy criterions. However, the accuracy on the validation data is 97% when using gini criterions. According to our results we adopted gini criterions to predict the test data. When we use trained model to predict test data, the accuracy attained 77%.

Question 5: Use the criterion selected above to train Decision Tree model on train data for min samples split=(2,5,10,20) and report the accuracies on the validation data. Select the best parameter and report the accuracy on the test data. (2 marks)

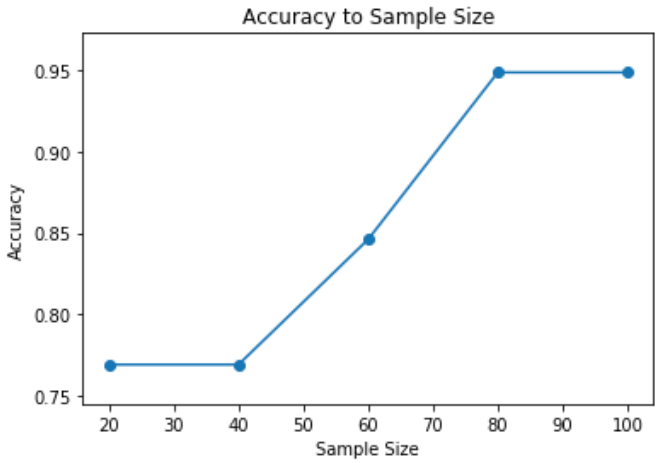
```
when min_sample_split is 2
we get the Validation accuracy = 0.897435897436
when min_sample_split is 5
we get the Validation accuracy = 0.974358974359
when min_sample_split is 10
we get the Validation accuracy = 0.948717948718
when min_sample_split is 20
we get the Validation accuracy = 0.923076923077

Test accuracy = 0.74358974359
```

When we use min_sample_split in 2,5,10, and 20,the accuracy on the validataion data is 90%, 97%, 95% and 92% respectively. Therefore, we select min_sample_split 5 as our parameters in our model. When we use trained model to predict test data, the accuracy attained 74%.

Question 6: Use the parameters selected above (Q4 and Q5) to train Decision Tree model using the first 20, 40, 60, 80 and 100 samples from train data. Keep the validation set unchanged during this analysis. Report and plot the accuracies on the validation data. (2 marks)

```
when our samples are first 20
We get validation accuracy = 0.769230769231
when our samples are first 40
We get validation accuracy = 0.769230769231
when our samples are first 60
We get validation accuracy = 0.846153846154
when our samples are first 80
We get validation accuracy = 0.948717948718
when our samples are first 100
We get validation accuracy = 0.948717948718
```



Question 7: Train k-nn model on train + validation data and report accuracy on test data. Use Euclidean distance and k=3. (1 mark)

```
Accuracy = 0.871794871795 at k = 3
```

Question 8: Train the model on train data for distance metrics dned by l1,linf , l2. Report the accuracies on the validation data. Select the best metric and report the accuracy on the test data for the selected metric. Use k=3. (1 mark)
Accuracy on the validation data

```
Accuracy = 0.948717948718 using distance metric manhattan
Accuracy = 0.923076923077 using distance metric chebyshev
Accuracy = 0.923076923077 using distance metric euclidean
```

Results showed us that the manhattan distance is te best metric for our validation data

Accuracy on the test data

```
Accuracy = 0.948717948718
```

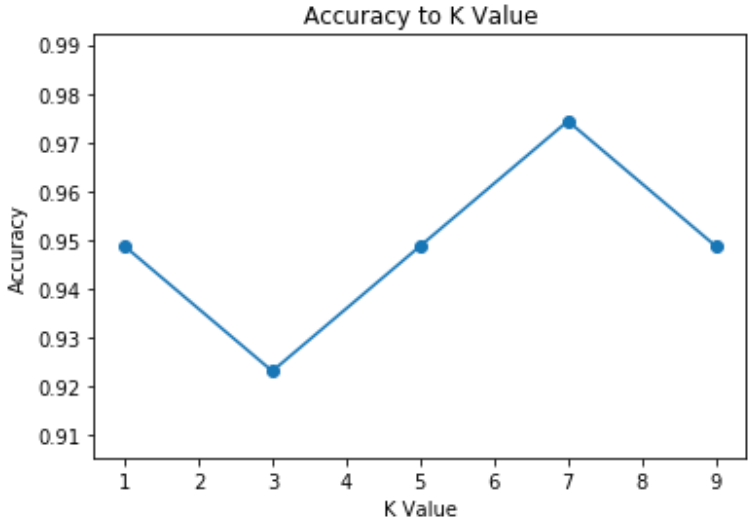
Question 9: Train the k-nn model on train data for k=1,3,5,7,9. Report and plot the accuracies on the validation data. Select the best 'k' value and report the accuracy on the test data for the selected 'k'. Use Euclidean distance. (2 marks)

Accuracy on the validation data

```
Accuracy = 0.948717948718 at k = 1
Accuracy = 0.923076923077 at k = 3
Accuracy = 0.948717948718 at k = 5
Accuracy = 0.974358974359 at k = 7
Accuracy = 0.948717948718 at k = 9
```

Results showed us when we use k = 7, we got the highest accuracy for our validation data
Accuracy on the test data

```
Accuracy = 0.948717948718 at k=7
```



Question 10: Instead of using full train data, train the model using the first 20, 40, 60, 80 and 100 data samples from train data. Keep the validation set unchanged during this analysis. Report and plot the accuracies on the validation data. Use Euclidean distance and k=3. Note: Don't shuffle the data and use only the first n samples', otherwise your answers may differ. (2 marks)

```
when our samples are 20:
We get validation accuracy = 0.948717948718
when our samples are 40: We get validation accuracy = 1.0
```

```
when our samples are 60: We get validation accuracy = 1.0
when our samples are 80: We get validation accuracy = 1.0
when our samples are 100: We get validation accuracy = 0.923076923077
```