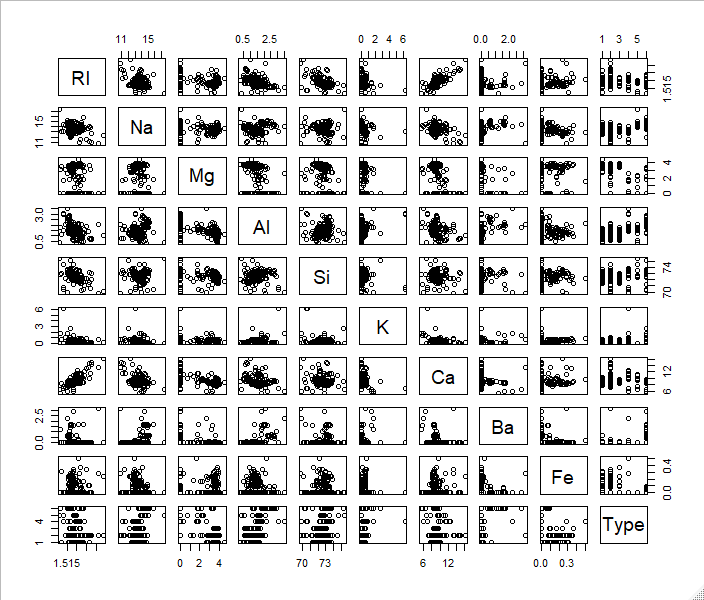
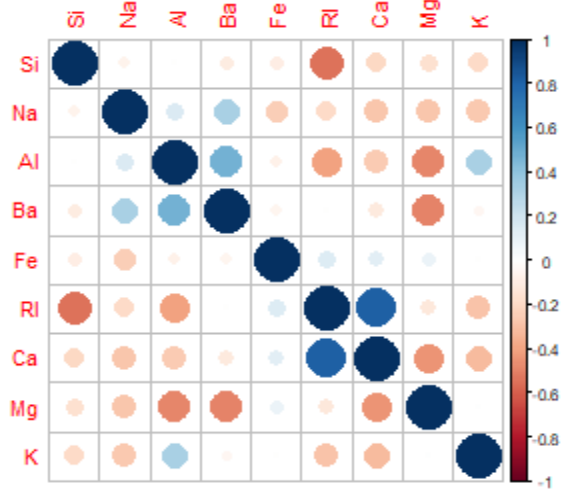
ASSIGNMENT 1

**3.1(a) Using visualizations, explore the predictor variables to understand their distributions as well as the relationships between predictors.**

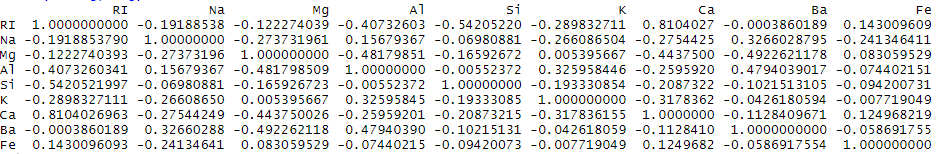
**PAIRWISE SCATTER PLOT:**



**CORRELATION STRUCTURE:**



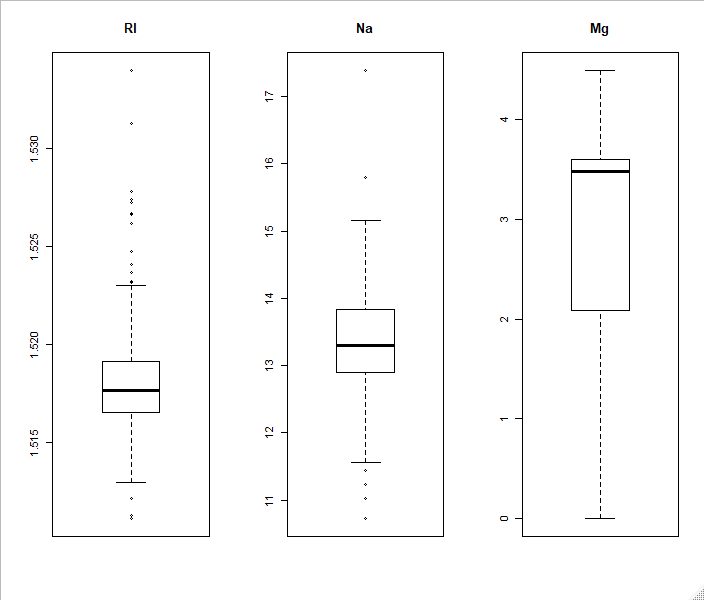
**CORRELATION:**

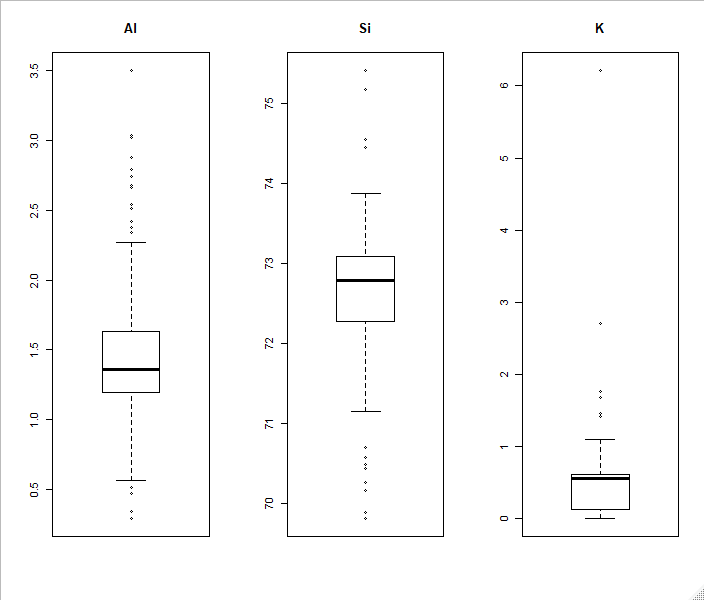
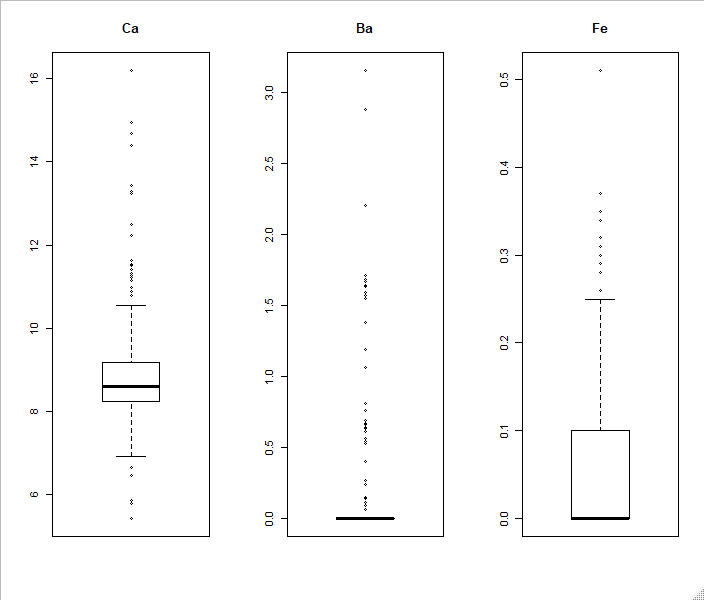


Correlation coefficient explains how dependent one variable is on the other. If the value of ‘R’ is high, then the two predictors are highly correlated and vice-versa irrespective of the positive or negative value. The positive value indicates that the predictors are directly proportional, whereas the negative value means they are indirectly proportional. From the above visualization we can infer that there exists significant correlation for Ri with Al, Si, Ca; Mg with Al, Ca.

**3.1(b) Do there appear to be any outliers in the data? Are any predictors skewed?**

**OUTLIERS:**

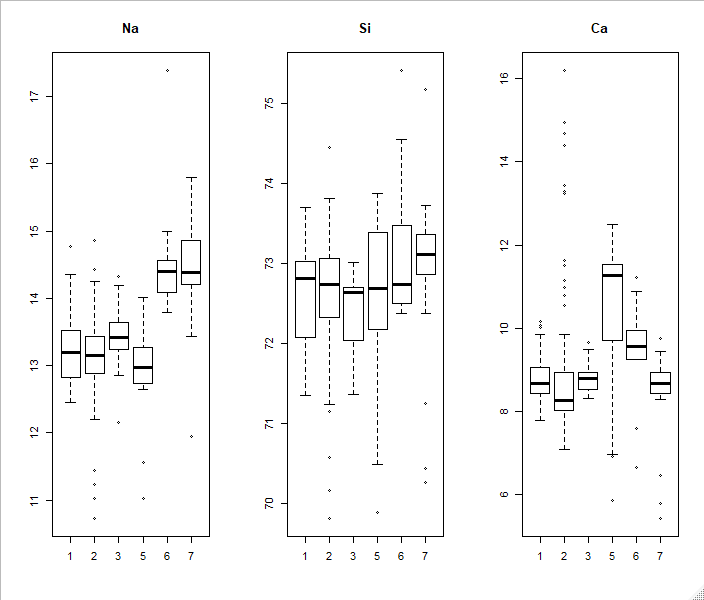


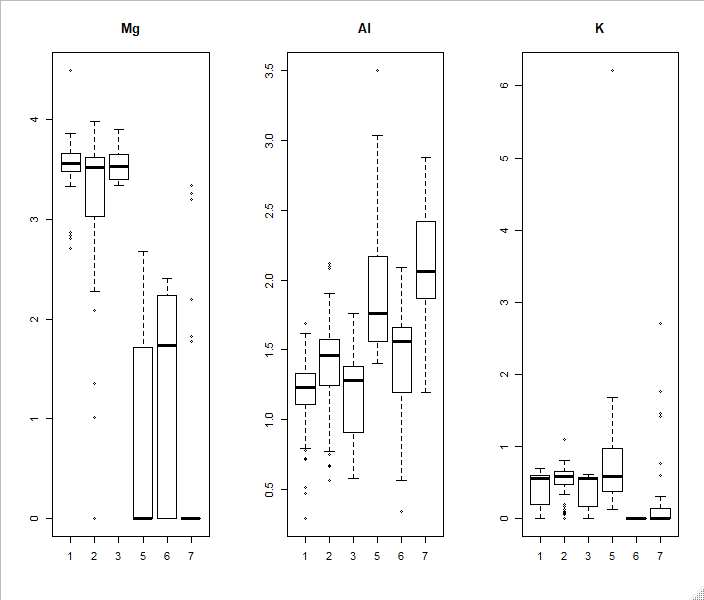


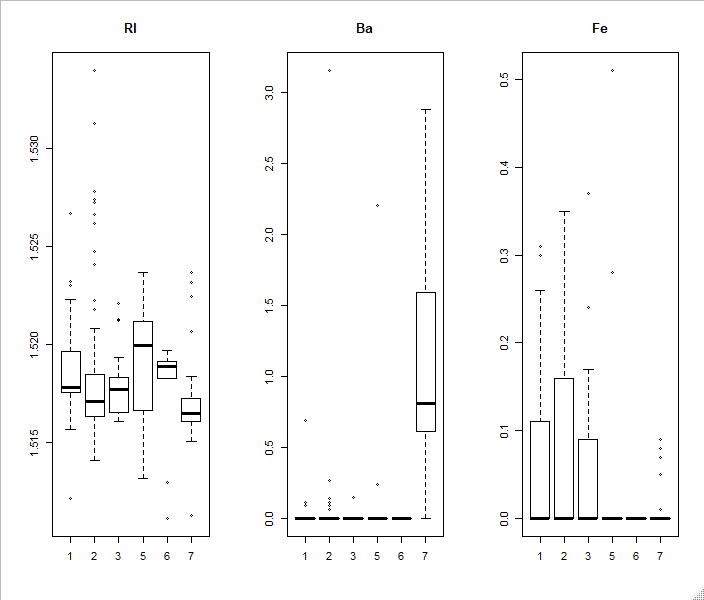
Data points lying beyond the whiskers are referred as outliers. From above visualizations, it can be inferred that Ri, Na, Al, Si, K, Ca, Ba, Fe have outliers. Only Mg do not have any outliers.

The distance of the ‘box’ (inter-quartile range) from the whiskers determines the skewness of data. It can be inferred that Ri, Na, Ca, Si, Al are symmetric, Mg is slightly skewed to the left and Fe, K, Ba are skewed towards right.

**VARIABLE vs TYPE:**



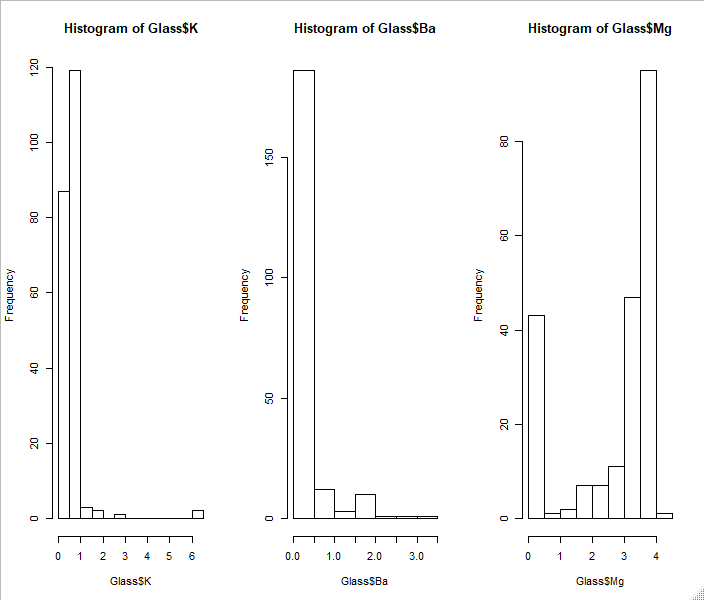


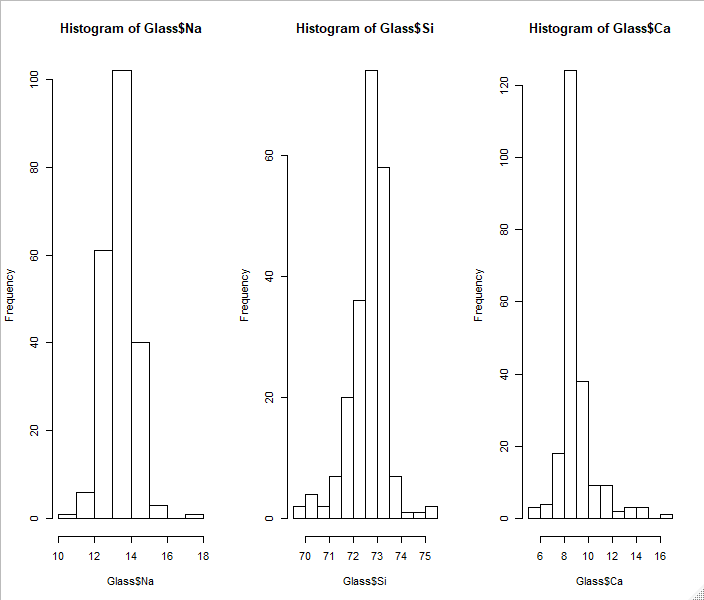


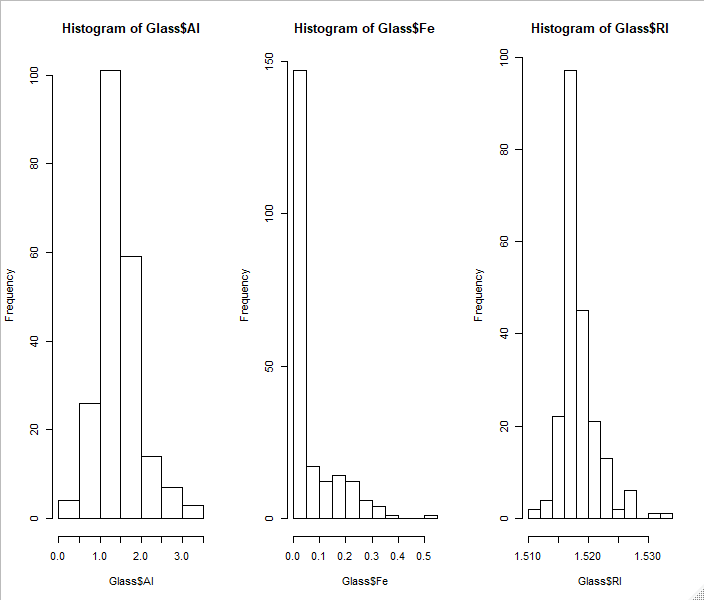
**SKEWNESS:**



**SKEWED PREDICTORS:**





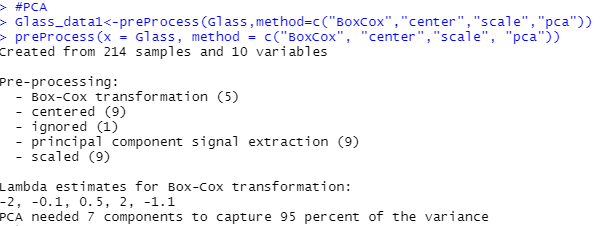


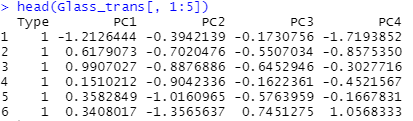
The above visualizations clearly indicate that Mg and Si are slightly left skewed, Ri, Na, Al, Fe are slightly right skewed, but K, Ba, Ca are extremely skewed towards right.

**(c) Are there any relevant transformations of one or more predictors that might improve the classification model?**

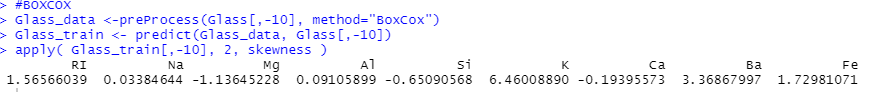
PCA and Box-cox transformations are performed to reduce the skewness of data.

**PCA:**



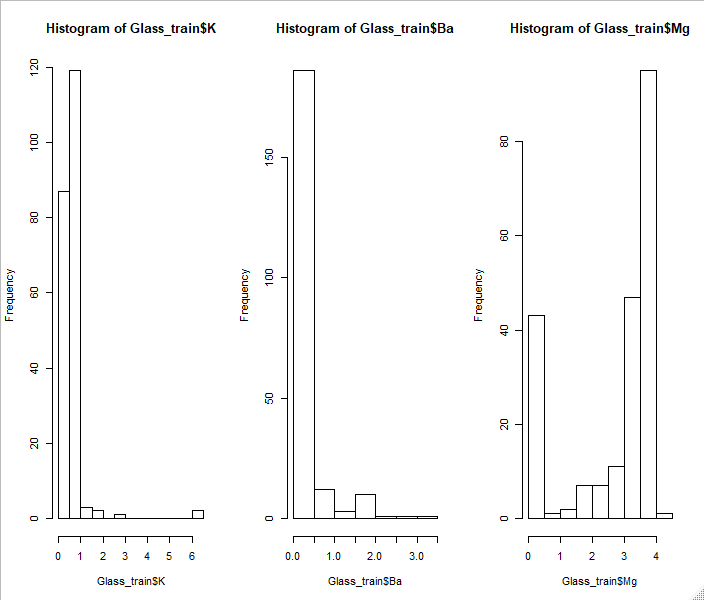


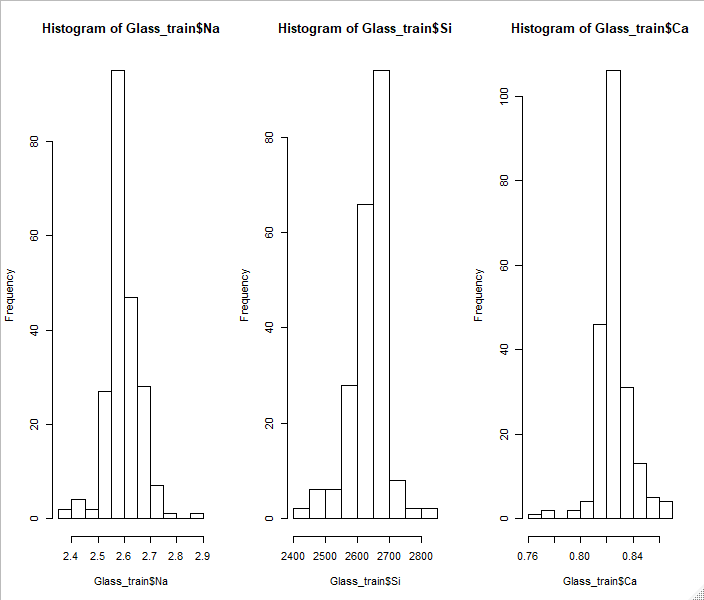
**BOXCOX TRANSFORMATION:**

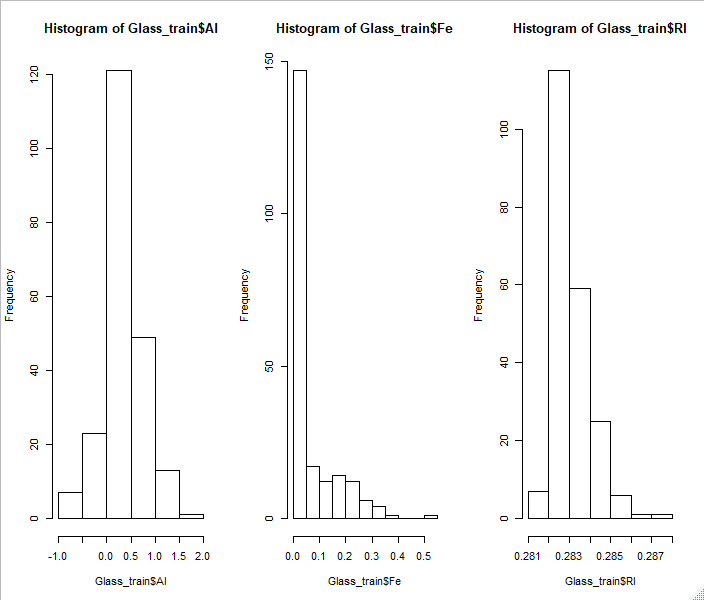


The histograms below depict the distribution of the data after transformations.

**TRANSFORMED SKEWED PREDICTORS:**



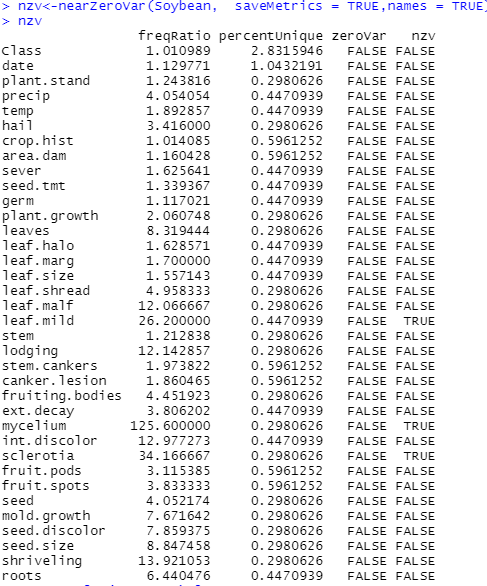




**3.2 (a) Investigate the frequency distributions for the categorical predictors. Are any of the distributions degenerate in the ways discussed earlier in this chapter?**

The distribution of data having very few distinct values or only one distinct value (all same values) is termed degenerate distribution. The ‘nearzerovar’ function determines all the variables having degenerate distribution.

**NEAR-ZEROVAR:**



From the above snapshot it is clear that, 3 predictors leaf.mild, mycelium, sclerotia have near zero variance, other predictors have highly varying values.

**3.2 (b) Roughly 18% of the data are missing. Are there particular predictors that are more likely to be missing? Is the pattern of missing data related to the classes?**

From the above bar plots, there are missing values in various predictors. The predictors plant.stand, precip, date, temp, hail, crop.hist, area.dam, sever, seed.tmt, germ, plant.growth, leaf.halo, leaf.marg, leaf.size, leaf.shread, leaf.malf, stem, leaf.mild, lodging, stem.cankers, ext.decay, mycelium, int.discolor, sclerotia, fruit.pods, fruit.spots, seed, mold.growth, seed.discolor, seed.size, shriveling, roots have missing values.

We can find the pattern of missing values by inspecting the values of each class.

A picture containing person

Description generated with high confidence

From above snapshot, it is clear that ‘2-4-d-injury’, ‘alternarialeaf-spot’, ‘anthracnose’, ‘bacterial-blight’, ‘bacterial-pustule’ have no missing values.

A screenshot of a cell phone

Description generated with high confidence

From the above snapshot it is clear that ‘brown-spot’, brown-stem-rot’, ‘charcoal’ has no missing values. The class ‘cyst-nematode’ has 24 missing values each in 14 rows, whereas ‘diaporthe-pod-&-stem-blight’ has 11 missing values in 9 rows and 13 missing values in 6 rows.

A screenshot of a cell phone

Description generated with very high confidence

From the above snapshot it is clear that ‘diaporthe-stem-canker’, ‘downy-mildew’, ‘frog-eye-leaf-spot’, ‘phyllosticta-leaf-spot’ has no missing values. The class ‘herbicide-injury’ has 20 missing values each in 8 rows whereas ‘phytophthora-rot’ has 13 missing values each in13 rows and 19 missing values in 56 rows.

A screenshot of a computer

Description generated with high confidence

From the above snapshot it is clear that ‘powdery-mildew’, purple-seed-stain’, rhizoctonia-root-rot’ has no missing values.

All the above snapshots clearly depict that, there are missing values only for the classes ‘cyst-nematode’, ‘diaporthe-pod-&-stem-blight’, ‘herbicide-injury’, ‘phytophthora-rot’ leading to existence of a pattern.