Multicollinearity is something that one faces on a regular basis in different datasets.

It can be dealt with multiple methods.

(1) **Correlation plot-** You can use corrplot function in R to draw a correlation plot and can easily set a threshold value to determine which two variables are showing multicollinearity. It can be done by using function corrplot.

(2) **Variance inflation factor(VIF)-** There are several packages in R that provide variance inflation factor function , generally package HH, vif in package car, and vif in package VIF all provide it.

A VIF for a single explanatory variable is obtained using the r-squared value of the regression of that variable against all other explanatory variables. Higher the value, the higher the collinearity.

We can eliminate the variables having high VIF values one by one and check consistently our accuracy because with removal of variable also refers to the loss of information.

(3) **Prinicipal component analysis(PCA)-** This method is generally used when we have large no of variables but it can be used in this case too, however first we need to convert the categorical variables into numeric by using one hot encoding

This method enables us to select few important variables that explain the maximum variability towards response variable.Generally the first principal component explains the maximum variability.

(4) **Partial least squares(PLS)-** This method also enables us to detect the important variables just like principal components and its can be used when there are more than one response variables.

We can use the plsdepot package for the same.

Apart from all these methods generally available algorithms can also be used like

Random forest - It has importance function in it and after applying random forest on a model like this its importance function will give us the importance values in terms of mean decrease Gini so we can remove variables with least value one by one and see improvement in performance.

Xgboost- It is one of most powerful algorithm as it has its own metrics which enables us to get the xgb.importance which has three metrics namely gaun, cover and frequency

Gain is the most important as it tells us the improvement in accuracy brought on by features added to it.

Decision trees- It is also one of the general basic algorithms but sometimes it works quite well for some datasets with good accuracy giving insight into important variables.

NOTE - Let's think of a scenario where we have a constant variable (all observations have same value, 7) in our data set. Will it explain variance of our response variable ,Of course not , because it has zero variance.

In case of high number of dimensions, we should drop variables having low variance compared to others because these variables will not explain the variation in target variables.