Pre Processing Steps:

1. Data Type conversions – Converting to numeric, integer, character, logical and factor types as per the requirement. Strings that look like dates should be converted to date/datetime data types in order to easily perform date manipulation or calculations. ***Examples – as.numeric(), as.character(), as.factor(), as.integer(). Use lubridate package for data related functions like mdy(), ymd() etc..***
2. String normalization – Involves trimming, creating substrings or removing special characters that are not required or characters that can cause unwanted results. ***Examples – grep(), gsub(), stringr package functions like str\_trim(), str\_sub(), str\_detect().***
3. Missing Values – Appropriately handling missing values as per the objective of the problem. For instance, using is.na = T argument for calculating descriptive statistics. In case of model building, imputation of missing values may be required (excluding some tree-based models). ***Examples – mean(x, is.na = TRUE), impute()***
4. Outliers – Presence of outliers can result in skewed distribution. Outliers should be identified and discussed during the preprocessing phase. ***Examples – boxplot()***
5. Standardization/Normalization – Center and scale the variables before building models. ***Examples – scale(x, center = TRUE, scale = TRUE)***
6. Collinearity – Variables with high degree of collinearity or Multicollinearity should be handled appropriately while building models. This sometime requires removing highly correlated variables before building models. ***Examples – cor(), corrplot(), cor.test()***