**Data Wrangling:**

This data set includes 569 observations of 33 variables. There are no missing values in the main 30 variables (features obtained from imaging data). The last column of the data set was removed as it included an unknown variable with NA values. Additionally, the column for patient ID was removed as it contained no useful information for our analysis. The column names were correct and there was no need for renaming.

'data.frame': 569 obs. of 33 variables:

$ id : int 842302 842517 84300903 84348301 84358402 …

$ diagnosis : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...

$ radius\_mean : num 18 20.6 19.7 11.4 20.3 ...

$ texture\_mean : num 10.4 17.8 21.2 20.4 14.3 ...

$ perimeter\_mean : num 122.8 132.9 130 77.6 135.1 ...

$ area\_mean : num 1001 1326 1203 386 1297 ...

$ smoothness\_mean : num 0.1184 0.0847 0.1096 0.1425 0.1003 ...

$ compactness\_mean : num 0.2776 0.0786 0.1599 0.2839 0.1328 ...

$ concavity\_mean : num 0.3001 0.0869 0.1974 0.2414 0.198 ...

$ concave.points\_mean : num 0.1471 0.0702 0.1279 0.1052 0.1043 ...

$ symmetry\_mean : num 0.242 0.181 0.207 0.26 0.181 ...

$ fractal\_dimension\_mean : num 0.0787 0.0567 0.06 0.0974 0.0588 ...

$ radius\_se : num 1.095 0.543 0.746 0.496 0.757 ...

$ texture\_se : num 0.905 0.734 0.787 1.156 0.781 ...

$ perimeter\_se : num 8.59 3.4 4.58 3.44 5.44 ...

$ area\_se : num 153.4 74.1 94 27.2 94.4 ...

$ smoothness\_se : num 0.0064 0.00522 0.00615 0.00911 0.01149 ...

$ compactness\_se : num 0.049 0.0131 0.0401 0.0746 0.0246 ...

$ concavity\_se : num 0.0537 0.0186 0.0383 0.0566 0.0569 ...

$ concave.points\_se : num 0.0159 0.0134 0.0206 0.0187 0.0188 ...

$ symmetry\_se : num 0.03 0.0139 0.0225 0.0596 0.0176 ...

$ fractal\_dimension\_se : num 0.00619 0.00353 0.00457 0.00921 0.00511 ...

$ radius\_worst : num 25.4 25 23.6 14.9 22.5 ...

$ texture\_worst : num 17.3 23.4 25.5 26.5 16.7 ...

$ perimeter\_worst : num 184.6 158.8 152.5 98.9 152.2 ...

$ area\_worst : num 2019 1956 1709 568 1575 ...

$ smoothness\_worst : num 0.162 0.124 0.144 0.21 0.137 ...

$ compactness\_worst : num 0.666 0.187 0.424 0.866 0.205 ...

$ concavity\_worst : num 0.712 0.242 0.45 0.687 0.4 ...

$ concave.points\_worst : num 0.265 0.186 0.243 0.258 0.163 ...

$ symmetry\_worst : num 0.46 0.275 0.361 0.664 0.236 ...

$ fractal\_dimension\_worst: num 0.1189 0.089 0.0876 0.173 0.0768 ...

$ X : logi NA NA NA NA NA NA ...

We created boxplots for each of the variables (here we illustrate radius\_mean as an example) to check for outliers. We separately checked the outliers in benign and malignant cases and identified a few outliers in each group. However, we decided to keep the outliers because of the inherent variability in human biological features and also the small size of our data set.

