In this case study, we want to investigate a red wine dataset, the red variant of the Portuguese ”Vinho Verde” wine. In particular, we try to find out the association between alcohol level and potential physio-chemical information. The predictors included in our full model are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates.

We first fit the full model that contains all predictors. The R^2 value is 0.67, meaning 67% of variation can be explained by the model. It is not very high, given most of the predictors are statistically significant due to the small p-values, except the free sulfur dioxide variable. Thus, we remove the variable and fit a reduced model to see whether the reduced model is adequate. By looking at the Anova table, we confirm that we can drop the free sulfur dioxide predictor. Even though all the predictors are statistically significant, the R^2 value does not change.

For illustration purposes, we will continue with the reduced model to check model assumptions. We first start with unusual observations: After calculating points with respect to leverage values, we find that around 7 percent of points are high leverage in which many of those are bad high leverage points because they do not follow the pattern of the rest of the data. The LS fitting would change a lot if we remove the points. We can studentized the residuals of the data to get the outliers, and there turns out to be no outliers as the residuals are all less than the critical value with Bonferroni correction. Last unusual observation is influential points, which we can detect by Cook’s distance. The maximum value of the Cook's distances is 0.072, so there's no high influential point.

Now, we proceed with checking constant variance, normality, and linearity assumptions by graphical and statistical tools. The constant variance assumption is violated: The p-value in the Breusch-Pagan test is less than 0.05, which rejects the null hypothesis of constant variance. This is also indicated in the plotted graph with a clear downward sloping trend of points with negative residuals. The normality assumption is violated: The p-value from the Kolmogorov-Smirnov test is smaller than 0.05, showing that the null hypothesis of normality is rejected, while there are departures on the tails of the distribution in the Q-Q plot. By looking at the partial regression plots, we noticed that the covariance of some predictors with the response are very small and the relationship is not linear, so the linearity assumption is violated.

In order to fix one or more of the assumptions, we considered the Box Cox transformation with 1 / Y^(-lambda). However, the unusual observations and violation of constant variance and normality still holds after transformation. As a result, other transformations need to be considered given the situation where we have low p-values and non-linearity.